Standards in Genomic Sciences (2013) **8:**124-164 DOI:10.4056/sigs.3457

The 15th Genomic Standards Consortium meeting

Lynn Schriml and Ilene Mizrachi (co-hosts, local organizers), Peter Sterk, Dawn Field, Lynette Hirschman, Tatiana Tatusova, Susanna Sansone, Jack Gilbert, David Schindel, Neil Davies, Chris Meyer, Folker Meyer, George Garrity, Lita Proctor, M.H. Medema, Yemin Lan, Anna Klindworth, Frank Oliver Glöckner, Tonia Korves, Antonia Gonzalez, Peter Dwayndt, Markus Göker, Anjette Johnston, Evangelos Pafilis, Susanne Schneider, K. Baker, Cynthia Parr, G. Sutton, H.H. Creasy, Nikos Kyrpides, K. Eric Wommack, Patricia L. Whetzel, Daniel Nasko, Hilmar Lapp, Takamoto Fujisawa, Adam M. Phillippy, Renzo Kottman, Judith A. Blake, Junhua Li, Elizabeth M. Glass, Lynn Schriml, Ilene Mizrachi, Petra ten Hoopen, Jack Gilbert, Rob Knight, Susan Holmes, Curtis Huttenhower, Steven L. Salzberg, Bing Ma, Owen White

Standards-enabled Research in Genomics

April 22-24th, 2013

Natcher Conference Center, NIH

April 22-23: GSC15 Plenary and Poster Sessions

April 24: GSC Working groups and GSC Hackathon

April 25-26: GSC15 Satellite Meeting: Genomic Observatories (Smithsonian)

Standards-enabled Research in Genomics

The 15th Genomic Standards Consortium (GSC15) meeting will be held at NIH (Bethesda, Maryland) from April 22-24th. This meeting will highlight the utilization of genome metadata standards for the enhancement of our biological understanding of microbes, the interaction between microbial genomes, human health and disease. GSC15 will provide a forum for standards developers, genome and metagenome researchers and biological database developers to discuss their research, initiate collaborations, join GSC working groups and engage in panel discussions. The conference will include two days of plenary talks featuring GSC projects and community standards efforts along with a keynote speaker, discussion of standards among a government panel and groups discussion panel. Day 3 of GSC15 will include concurrent GSC working groups **open to GSC15 participants**.

Meeting Logistics

http://www.nih.gov/about/visitor/index.htm

All visitors without an NIH ID badge will need to go through a short inspection at the Gateway Center, Bldg 66 area. All bags are inspected and scanned. Please present a government issued ID, driver's license or passport. If your passport is from Syria, Cuba, Iran, or Sudan, you are required to obtain approval 3 weeks prior to the meeting. Contact admin@ncbi.nlm.nih.gov for guidance.

Travel

There are three airports in the area:

Ronald Reagan National Airport (DCA) in Virginia
Dulles International Airport (IAD) in Virginia

Baltimore Washington International Airport (BWI) in Maryland

Hotels

A number of hotels are conveniently located near the NIH meeting location [Figure 1-5].

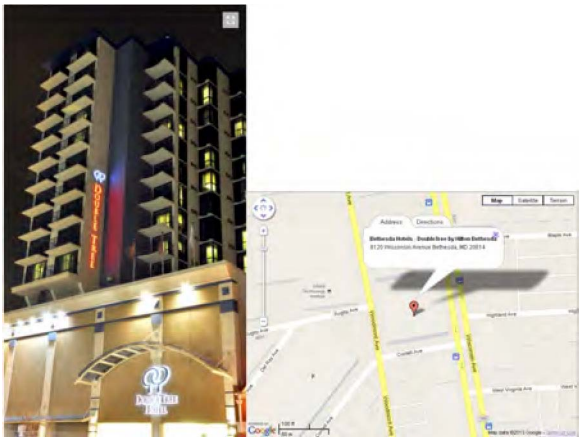


Figure 1. Double Tree Hotel – Walking distance to NIH, 8120 Wisconsin Avenue, Bethesda, MD 20814, Tel: 301-652-2000 /1-800-222-8733, http://www.doubletreebethesda.com

Figure 2. Hilton Garden Inn in Bethesda, 7301 Waverly Street, Bethesda, MD 20814, Tel: 1-301-654-8111, Fax 1-301-654-6584, http://hiltongardeninn.hilton.com, Walking distance to Bethesda Metro Station. One metro stop south of Medical Center/NIH

Figure 3. Hyatt Regency Bethesda, 1 Bethesda Metro Center, Bethesda, MD 20814, Tel: 657-1234, http://www.bethesda.hyatt.com, Walking distance to Metro Station. One metro stop south of Medical Center/NIH

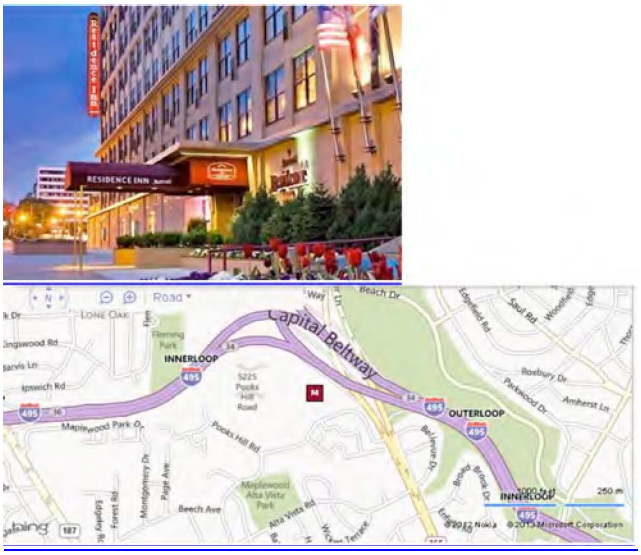


Figure 4. Residence Inn by Marriott, 7335 Wisconsin Avenue, Bethesda, MD 20814, Tel: 301-718-0200, http://www.residenceinnbethesdahotel.com/ or http://cwp.marriott.com/wasbt/nih/

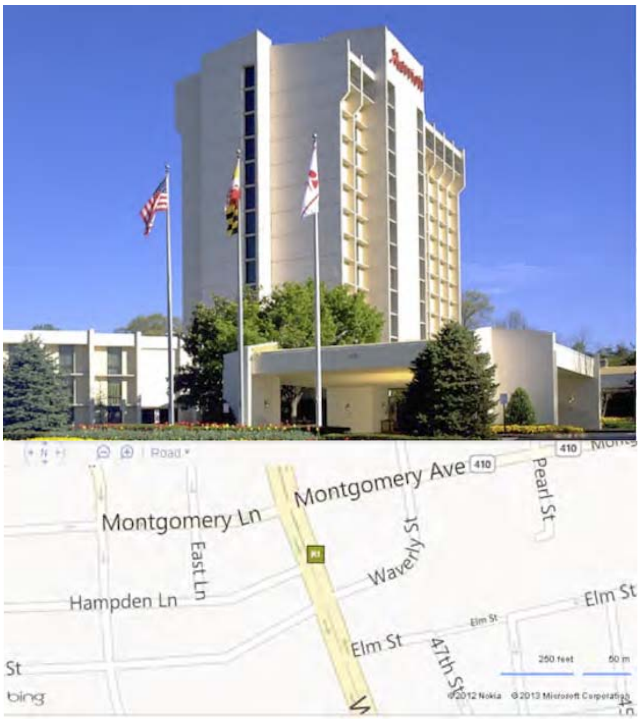


Figure 5. Marriott Pooks Hill, 5151 Pooks Hill Road, Bethesda, Md 20814, Tel: 301-897-9400; 1-800-228-9290, http://www.marriott.com/hotels/travel/wasbt-bethesda-marriott/, Shuttle from hotel to Medical Center

Agenda

Monday, April 22nd

8:00-9:00 Arrival, Registration, Poster setup, Coffee & Breakfast

9:00-9:15 Meeting Opening: Lynn Schriml and Ilene Mizrachi

Jim Ostell of NCBI opens GSC15

9:15-9:45 Standards in the INSDC

Ilene Mizrachi (NCBI) and Petra Ten Hoopen (EBI)

9:45-10:00 GSC Introduction

Dawn Field (NERC Centre for Ecology and Hydrology, UK)

10:00-10:30 GSC Groups I: Success stories of GSC standards in Environmental Genomics

Session Chair: Dawn Field

Konstantinos Liolios (DOE Joint Genomes Institute, USA)

The Genomes OnLine Database (GOLD) v.5: status of genomic and metagenomic projects and their associated metadata, GSC15p20

Jack Gilbert (Argonne National Laboratory and University of Chicago, USA)

The Hospital Microbiome Project – understanding microbial transmission and ecology in a health care environment, GSC15p38

Frank Oliver Glöckner (Max Planck Institute Bremen and Jacobs University Bremen, Germany)

Micro B3: Marine Microbial Biodiversity, Bioinformatics, Biotechnology, GSC15p05

10:30-11:00 Coffee Break & Poster Session

11:00-12:30 Standards in the HMP and Human Microbiome Research

Session Chairs: Rob Knight and Owen White

Owen White (University of Maryland School of Medicine, USA)

Utilization of Standardized Vocabularies in the Management of Human Microbiome Data, GSC15p43

Rob Knight (University of Colorado Boulder, USA)

Using standards-compliant metadata to combine human microbiome studies, GSC15p32

Curtis Huttenhower (Harvard School of Public Health, USA)

Cataloging genes and pathways in the human microbiome, GSC15p40

Jack Gilbert (Argonne National Laboratory & University of Chicago, USA)

The Earth Microbiome Project and the importance of data standards in modeling the planets microbiome, GSC15p37

Susan Holmes (Stanford University, USA)

Reproducible research for data normalization, analyses and simulations, GSC15p39

12:30-13:30 Lunch

13:30-14:15 Introduction: Owen White]

Keynote Speaker: Steven Salzberg

(Johns Hopkins University School of Medicine, USA)

14:15-15:00 **GSC Groups II**: GSC Standards ImplementationSession Chair: [Frank Oliver Glöckner]

Renzo Kottmann (Max Planck Institute Bremen, Germany)

Earth Sampling App for Ocean Sampling Day, GSC15p27

George Garrity (Michigan State University, USA)

Standards in Genomic Sciences (SIGS), GSC15p34

Antonio Gonzalez (University of Colorado Boulder, USA)

Next generation software pipelines and reproducibility for microbial community analysis, GSC15p07

Pelin Yilmaz (Max Planck Institute Bremen, Germany)

Phylogeny, ecology and biogeography of marine bacterial and archaeal clades, GSC15p33

15:00-15:30 Coffee Break & Poster Session

15:30-17:00 GSC Outreach I: Standards Development and Implementation in Genomic and Metagenomic Research (Submitted Abstracts) Session Chair: [Susanna-Assunta Sansone]

Anjantte Johnston (National Center for Biotechnology Information, USA)

GSC-Compliance at NCBI, GSC15p11

Heather Huot-Creasy (University of Maryland School of Medicine, USA)

The Human Microbiome Project Metadata Catalog, GSC15p19

Tonia Korves (The MITRE Corporation, USA)

Applying Publicly Available Genomic Metadata to Disease Outbreak Investigation, GSC15p06

Adam Phillippy (University of Maryland, USA)

Sequencing and analysis standards for microbial forensics, GSC15p26

Junhua Li (BGI, China)

A new version of human gut microbial gene catalog integrating with 3 continent populations and human gut-related bacterial genomes, GSC15p30

17:00-17:30 Poster Session Flash talks (2 min, 1 slide)

18:00 Day 1 Wrap Up

18:30-21:00- GSC15 Reception

Evening reception, food and drink provided by the GSC.

Lebanese Taverna Restaurant

7141 Arlington Road, Bethesda, MD

Tuesday, April 23rd

8:00-9:00 Coffee & Breakfast

9:00-10:30 GSC Outreach II

Standards Development and Implementation in Genomic and Metagenomic Databases and Applications (Submitted Abstracts)Session Chair: [Norman Morrison]

Judith Blake (The Jackson Laboratory, USA) & **Suzanna Lewis** (Lawrence Berkeley National Laboratory, USA)

Gene Ontology: Functional Annotation For Comparative Genomics, GSC15p29

Evangelos Pafilis (Hellenic Center of Marine Research, Greece)

ENVIRONMENTS: identification of environment descriptive terms in text, GSC15p12

Susanna-Assunta Sansone (University of Oxford, UK)

The ISA Commons: curate, visualize, analyze, share and publish, GSC15p15

Trish Whetzel (BMIR Stanford University, USA)

NCBO Technology: Powering Semantically Aware Applications to

enable Standards Driven Research, GSC15p22

K. Eric Wommack (University of Delaware, USA)

The VIROME Compare-inator: a web tool for exploring the diversity of viral communities, GSC15p21

Hilmar Lapp (National Evolutionary Synthesis Center, USA)

The Blessing and the curse: handshaking between general and specialist data repositories

10:00-10:30 Coffee Break & Poster Session

10:30-12:00 GSC Outreach III

Standards Development and New Standards in Genomic and Metagenomic Databases and Applications (Submitted Abstracts) Session Chair: Linda Amaral-Zettler

Granger Sutton (J. Craig Venter Institute, USA)

Standards for Pan-Genome Data Repositories, GSC15p18

Folker Meyer (Argonne National Laboratory and University of Chicago, USA) & *Tatiana Tatusova* (National Center for Biotechnology Information, USA)

Pan-genome: a new paradigm in microbiology, GSC15p28

Cynthia Parr (Smithsonian Institution, USA)

Encyclopedia of Life: Applying concepts from Amazon.com and LEGO to biodiversity informatics, GSC15p17

Katharine Barker (Smithsonian Institution, USA)

The Global Genome Biodiversity Network (GGBN) and the Global Genome Initiative (GGI): Building the Infrastructure for the future of Genomics, GSC15p16

Marnix Medema (Max Planck Institute Bremen, Germany)

MIBiG: Minimal Information about a Biosynthetic Gene Cluster, GSC15p02

Markus Göker (DSMZ, Germany)

Proposal for a Minimum Information on a Phenotype MicroArray Study (MIPS) standard, GSC15p09

Michael Schneider (Max Planck Institute Bremen, Germany)

A MIxS and DwC compliant biodiversity Reference Data Model to support the community in implementing the database layer, GSC15p14

12:00-13:30 Lunch

13:30- 15:30 Government Panel

Discussion Topic: Where are standards in genomics most needed now?Session Chairs: Lynette Hirschman and Lynn Schriml

Panel Chair: Dan Drell (DOE)

Session Members: Susan Gregurick (DOE), Adam Phillippy (DHS), Alison Yao (NIAID), Ann Lichens-Park (NIFA/USDA), Lita Proctor (Common Fund), Marc Allard (FDA), Marc Salit (NIST), Patricia Reichelderfer (NICHD), Vivien Bonazzi (NHGRI), Sylvia Spengler (NSF/CISE)

15:30- 16:00 Coffee Break & Poster Session

16:00-17:00 Panel Discussion

Proposed topics

1. The need for new standards: PanGenomes

2. Genome Sequence Annotation Standards

17:00 Day 2 Wrap Up. Close of GSC15 plenary sessions.

Official Handoff of GSC15 to GSC16 to be held in Brisbane, Australia

Wednesday, April 24th

Working group sessions: Open to All GSC15 Participants

8:00	Breakfast & Coffee
8:30-12:00	Morning Session
	Morning Session: Genomic Observatories (GOs) Session [Balcony B]
	(all groups in GOs session), Program: See http://tinyurl.com/axh44uq
10:15-10:45	Morning Coffee Break
12:00-1:00	Lunch
13:00-15:00	Afternoon Concurrent Session I:
	GOs core group meeting [Balcony B]
	M5 Working Group [Room H] GSC Biodiversity Hackathon & Compliance and Interoperability Working Group [Room C1/C2]
14:45-15:15	Afternoon Coffee Break
15:15-17:00	Afternoon Concurrent Session II
	GOs core group meeting [Balcony B]
	M5 Working Group [Room H] GSC Biodiversity Hackathon & Compliance and Interoperability Working Group [Room C1/C2]
	GSC Board Meeting [Closed Session] [Room B]

April 25-26th

GSC15 Satellite Meeting: Genomic Observatories 2nd

(GOs2) International Meeting (Smithsonian)

Please see Genomic Observatories Network workshop (GOs2) for details of the meeting, and http://genomicobservatories.org for more information about the Genomic Observatories Network in general.

GSC15 Poster presenters

Poster Number, Presenter/Authors, Title

1. ***Lita Proctor***

   The trans-NIH Microbiome Working Group (TMWG), GSC15p01

2. ***Yemin Lan***, Nivedita Clark, Christopher Blackwood and Gail Rosen

   Identifying functional signatures in microorganism genomes related to polymer decomposition, GSC15p03

3. ***Michael Vyverman***, Bernard De Baets, Veerle Fack and ***Peter Dawyndt***

   ALFALFA: a Novel Algorithm for Long Fragment Mapping and Alignment, GSC15p10

4. Hubert Denise, ***Peter Sterk***, Matt Corbett, Matthew Fraser, Craig McAnulla, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Maxim Scheremetjew and Sarah Hunter

   The Metagenomics Resource at the European Bioinformatics Institute, GSC15p13

5. Daniel Nasko, Jaysheel Bhavsar, Shawn Polson and ***K. Eric Wommack***

   Assessing patterns of CRISPR spacer composition using the CASC discovery and validation tool, GSC15p21

6. ***Takatomo Fujisawa***, Toshiaki Katayama, Shinobu Okamoto, Hiroshi Mori, Yasunori Yamamoto and Yasukazu Nakamura

   Development of an ontology for the INSDC Feature Table Definition, GSC15p01

7. Adam Arkin, Robert Cottingham, Sergei Maslov, Rick Stevens, Thomas Brettin, Dylan Chivian, Paramvir Dehal, Christopher Henry, Folker Meyer, Jennifer Salazar, Doreen Ware, David Weston, Brian Davison and ***Elizabeth M. Glass***

   Using Data Standards in KBase: The DOE Systems Biology Knowledgebase,. GSC15p31

8. ***Elizabeth M. Glass***, Yekaterina Dribinsky, Pelin Yilmaz, Hal Levin, Robert Van Pelt, Doug Wendel, Andreas Wilke, Jonathan Eisen, Sue Huse, Anna Shipanova, Mitch Sogin, Jason Stajich, Rob Knight, Folker Meyer and Lynn Schriml

   Meta Data Standards for the Built Environment - MIxS-BE, GSC15p32

9. Anna Klindworth, Elmar Pruesse, Timmy Schweer, Joerg Pelpies, Christian Quast and ***Frank Oliver Glöckner***

   In silico evaluation of primer and primer pairs for 16S ribosomal RNA biodiversity, GSC15p04

10. Bart Mesuere, Bart Devreese, Griet Debyser, Maarten Aerts, Peter Vandamme and ***Peter Dawyndt***

    Unipept: Exploring biodiversity of complex metaproteome samples, GSC15p08

11. ***Bing Ma***, Arthur Brady, Owen White and Jacques Ravel

    Organism-centric approach to study paired metagenome and metatranscriptome profiles of vaginal microbial community, GSC15p42

12. ***Peter Sterk***: GSC Projects

13. NCBI

14. Anjanette Johnston, John Anderson, Tanya Barrett and Ilene Mizrachi

    GSC-Compliance at NCBI, GSC15p11

15. Evangelos Pafilis, Sune Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Katerina Vasileiadou, Christos Arvanitidis and Lars Juhl Jensen

    ENVIRONMENTS: identification of environment descriptive terms in text, GSC15p12

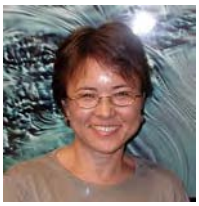
16. ***Trish Whetzel***

    NCBO Technology: Powering Semantically Aware Applications to enable Standards Driven Research, GSC15p22

17. Marnix Medema, Pelin Yilmaz, Renzo Kottmann, Frank Oliver Glöckner, Rainer Breitling and Eriko Takano

    MIBiG: Minimal Information about a Biosynthetic Gene Cluster, GSC15p02

GSC15 Abstracts



Lita Proctor

GSC15p01
The trans-NIH Microbiome Working Group (TMWG)

Lita Proctor

National Institutes of Health, Bethesda, MD, USA

Correspondence: lita.proctor@nih.gov

Keywords: microbiome, trans-NIH, working group

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

In recent years, interest in the microbiome has greatly increased. A newly established group of NIH extramural program staff representing 14 Institutes and Centers (ICs) with a common interest in the microbiome formed in response to this growing interest. NIH is a large agency and even internally, it's not always possible to know what areas each IC is funding or planning to fund. The trans-NIH Microbiome Working Group (TMWG) formed to 1) create an internal forum for exchanging information about current microbiome-related funding opportunities, plan joint microbiome-related funding opportunities and explore possible joint microbiome-related activities with other agencies but also 2) establish a clearinghouse for communicating NIH microbiome-related news to the broader scientific community. This poster will highlight some of the current and upcoming activities of the TMWG.

M.H. Medema

GSC15p02
MIBiG: Minimal Information about a Biosynthetic Gene Cluster

M.H. Medema[1,2], P. Yilmaz[3], R. Kottmann[3], F.O. Glöckner[3], R. Breitling[2,4], E. Takano[1,4]

[1]Department of Microbial Physiology

[2]Groningen Bioinformatics Centre, University of Groningen, Groningen, The Netherlands

[3]Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

[4]Faculty of Life Sciences, Manchester Institute of Biotechnology, University of Manchester, United Kingdom

Correspondence: m.h.medema@rug.n

Keywords: genomic standards, biosynthetic gene cluster, secondary metabolism, natural products

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

Bacteria, fungi and plants produce an enormous variety of secondary metabolites with manifold biological activities, e.g., as antibiotics, immunosuppressants, and signaling molecules. The biosynthesis of such molecules is encoded by compact genomic units: biosynthetic gene clusters. Over the past decades, hundreds of biosynthetic gene clusters encoding the biosynthesis of secondary metabolites have been characterized. Although dozens of biosynthetic gene clusters are published and thousands are sequenced annually (with or without their surrounding genome sequence), very little effort has been put into structuring this information. Hence, it is currently very difficult to prioritize gene clusters for experimental characterization, to identify the fundamental architectural principles of biosynthetic gene clusters, to understand which ecological parameters drive their evolution, and to obtain an informative 'parts registry' of building blocks for the synthetic biology of secondary metabolite biosynthesis. Therefore, developing a genomic standard for experimentally characterized biosynthetic gene clusters would be of great value. The standard will build on the MIxS standards for ecological and environmental contextualization [1]. Additionally, biochemical, genomic and pharmaceutical metadata will be added in parameters such as enzyme substrate specificities, operon structure, chemical moieties of the product, metabolic precursors and compound bioactivity. Using the already developed computational pipeline for gene cluster analysis antiSMASH [2], information on characterized biosynthetic gene clusters will be linked to the untapped wealth of thousands of unknown gene clusters that have recently been unearthed by massive genome sequencing efforts. Taken together, this has the potential to guide the characterization of new metabolites by allowing to optimize the sampling of diversity at different levels and to identify the biochemical, genomic and ecological parameters that are key predictors of pharmaceutically relevant biological activities. Moreover, it can transform the unordered pile of literature on secondary metabolites into a structured and annotated catalogue of parts that can be used as building blocks to design new biochemical pathways with synthetic biology [3,4].

Yemin Lan

GSC15p03
Identifying functional signatures in microorganism genomes related to polymer decomposition

Yemin Lan[1], Nivedita Clark[2], Christopher Blackwood[2] and Gail Rosen[1].
[1]Drexel University
[2]Kent State University
Correspondence: yeminlan@gmail.com
Keywords: polymer decomposition, functional signatures, feature selection
Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

Plant cell wall polymers, such as cellulose, xylan, pectin (PGA) and lignin, must be degraded into monomers before being taken up and metabolized by microorganisms. The degradation of polymers, however, is likely to be different among microorganisms, and with growth on different polymers. This variation has been poorly characterized, but could allow for a variety of investigations that help us interpret the ecological strategy microorganisms or microbial communities decompose plant material.

By monitoring the respiration rate of 15 microorganisms cultured on four different polymer substrates, and analyzing the gene content of their complete genomes, our project goal is to identify functional signatures of various categories that characterize polymer degradation ability of microorganisms, cultured on different substrates.

The complete genomes of 15 soil microorganisms are available from NCBI, whose genes can be mapped to various functional categories, including Metacyc Pathways, Pfams (protein family), Gene Ontology terms and eggNOG orthologous groups. Feature selection methods (TF-iDF and mRMR) and Pearson correlation are used to identify a selected set of functional genomic signatures that best distinguish microorganisms that degrade substrate slowly from those that degrade substrate quickly. We then use support vector machine to classify the genomes based on abundances of selected pathways/Pfams/GO terms/eggNOGs within each genome, and principal component analysis for ordination. We show that better classification or better ordination was achieved when using selected signatures for some cases with PGA and xylan as substrates (with area under the receiver operating characteristic curve 10-15% higher than chance), but not for the others. The method shows the potential of revealing the underlying functional mechanisms important in determining the ability of microorganisms to decompose different polymers.

Anna Klindworth

GSC15p04
In silico evaluation of primer and primer pairs for 16S ribosomal RNA biodiversity

Anna Klindworth[1], Elmar Pruesse[2], Timmy Schweer[1], Joerg Pelpies[3], Christian Quast[1] and Frank Oliver Glöckner[1]

[1]Max Planck Institute for Marine Microbiology, Bremen, Germany

[2]Jacobs University, Bremen, Germany

[3]Ribocon, Bremen, Germany

Correspondence: fog@mpi-bremen.de

Keywords: 16S rDNA, primer, next generation sequencing, diversity analysis, SILVA TestPrime

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

16S ribosomal RNA gene (rDNA) amplicon analysis remains the standard approach for the cultivation-independent investigation of microbial diversity. However, the accuracy of these analyses depends strongly on the choice of primers. This issue has been addressed by an *in silico* evaluation of primer with respect to the SILVA non-redundant reference database (SILVA SSURef NR). A total of 175 primers and 512 primer pairs were analyzed with respect to overall coverage and phylum spectrum for Archaea and Bacteria. Based on this evaluation a selection of 'best available' primer pairs addressing different sequencing platforms is provided in order to serve as a guideline for finding the most suitable primer pair for 16S rDNA analysis in any habitat and for individual research questions. Moreover, the SILVA team developed a new SILVA TestPrime tool (http://www.arb-silva.de/search/testprime) allowing the scientific community to perform an online *in silico* PCR with their primer pair of interest. Re-evaluation using an up-to-date database can be assured and evaluation of primer pairs prior to amplification both remain attractive in the future [5,6].

Frank Oliver Glöckner

GSC15p05
Micro B3: Marine Microbial Biodiversity, Bioinformatics, Biotechnology

Frank Oliver Glöckner

Jacobs University Bremen, Bremen, Germany

Correspondence: fog@mpi-bremen.de

Keywords: Biodiversity, Bioinformatics, Biotechnology, Standards, Ocean Sampling Day

Program session: Development of Resources, Tools or Databases Related to the GSC Mission, Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

The 32 partner Ocean of Tomorrow Project Micro B3 (Biodiversity, Bioinformatics, Biotechnology, www.microb3.eu) forms teams of experts in bioinformatics, computer science, biology, ecology, oceanography, bioprospecting, biotechnology, ethics and law. The consortium's main aims are to bring together the existing bodies of expertise in ecosystems biology, the processing and interpretation of data, modelling and prediction and the development of intellectual property agreements for the exploitation of high potential commercial applications. At its core Micro B3 aims to develop an innovative, transparent and user friendly open-access system, which will allow for seamless processing, integration, visualisation and accessibility of the huge amount of data collected in ongoing sample campaigns and long-term observations. This will in turn offer new perspectives for the modelling and exploration of marine microbial communities for biotechnological applications.

A key boost to the work will be provided by the Ocean Sampling Day (OSD, www.oceansamplingday.org), scheduled to take place on summer solstice - 21 June 2014. OSD will take place worldwide, with pilots conducted in 2012 and 13 to establish standardized sampling techniques. Adhering to the Minimum information checklists (MIxS) standard for describing molecular samples as outlined by the Genomic Standards Consortium will be essential for OSD. The event will generate a massive amount of useful marine microbial data to be included in the project's integrated MB3-Information System, providing the members of the biotechnology team with information to generate hypotheses for more cost- and time-efficient biotechnological testing and applications.

In summary Micro B3 is set to revolutionise Europe's capacity for bioinformatics and marine microbial data integration, to the benefit of a variety of disciplines in bioscience, technology, computing, standardisation and law.

Micro B3 is financially supported by the 7FP Ocean of Tomorrow Grant #287589

Tonia Korves

GSC15p06
Applying Publicly Available Genomic Metadata to Disease Outbreak Investigation

Tonia Korves, Matthew Peterson, Wenling Chang and Lynette Hirschman

MITRE, Bedford, MA, USA and McLean, VA, USA

Correspondence: tkorves@mitre.org

Keywords: disease outbreaks, data integration, public databases, applying genomic metadata

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

An important application of genomic metadata is in the investigation of disease outbreaks. The ability to identify the sources of disease outbreaks can prevent repeat events and potentially curtail outbreaks, saving lives, preventing societal disruption, and reducing economic costs. One way to discover the source of a pathogen is to identify other strains with shared biological properties, and then

use associated metadata to discover and evaluate potential sources. This process relies on metadata such as collection date, isolation source, and geographic location, and depends on the extent and format in which data is captured. Currently, this data is captured in a variety of formats and information sources, making it challenging to use for pathogen source identification.

We will present our work on methods for assembling and integrating metadata from public sources for source identification. We will illustrate this with a proof-of-concept, presenting a mock outbreak investigation of a *Salmonella enterica* strain. This investigation will start with information about an outbreak strain's DNA sequence and near phylogenetic relatives, and evaluate how public information sources can be used to address where and what type of environment the strain might have come from, whether the strain is an accidental laboratory escapee, and what laboratories have related strains needed for evaluating candidate origins. In the mock investigation, we will utilize a PostgreSQL database that we designed for metadata needed for source investigations, tools we created for automated importation and parsing of metadata from NCBI's BioProject and BioSample, and the LabKey platform to integrate, query, and present data from multiple sources. Data sources will include NCBI databases, MedLine, an MLST database, and StrainInfo. We will discuss the extent to which these tools and current publicly available data can address pathogen origin questions, and insights this might provide for standards and data capture efforts.

Antonio Gonzalez

GSC15p07
Next generation software pipelines and reproducibility for microbial community analysis

Antonio Gonzalez[1], Yoshiki Vázquez-Baeza[2], Will Van Treuren[2], Meg Pirrung[3], J. Gregory Caporaso[4,5] and Rob Knight[3]

[1]BioFrontiers Institute, University of Colorado, Boulder, CO, USA

[2]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO, USA

[3]University of Colorado Denver, Anschutz Medical Campus, Denver, CO, USA

[4]Department of Computer Science, Northern Arizona University, Flagstaff, AZ, USA

[5]Institute for Genomics and Systems Biology, Argonne National Laboratory, Argonne, IL, USA

Correspondence: antgonza@gmail.com

Kewords: software reproducibility, next generation software tools, microbial analysis

Program session: Development of Resources, Tools or Databases Related to the GSC Mission, Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

New sequencing technologies both produced an explosion of data and inspired a proliferation of individual scripts written by scientists rather than software developers. These ad hoc scripts were typically not based in software development techniques, and lead to a crossroad comparable to the "software crisis" of the 1970s; projects running over budget, overtime, inefficient and hard-to-maintain software, etc. Here we present the use case of the development of QIIME (Quantitative Insights Into Microbial Ecology), which is based on test-driven and agile software development techniques. Test-driven development is the concept of creating positive and negative controls for software, resulting in more robust systems and avoid-

ing common errors. Agile development is a methodology that allows adaptive planning in a collaborative environment, which provides a suitable environment for the creation of bioinformatics' tools. These methodologies not only ensure the reproducibility of results from the software but also facilitate rapid development. We also present Evident, next-generation software for microbial ecology, which runs within a browser and allows researchers to define the sampling effort for new studies by comparing to and relying on results from previously published datasets.

Peter Dwayndt

GSC15p08
Unipept: Exploring biodiversity of complex metaproteome samples

Bart Mesuere[1], Bart Devreese[2], Griet Debyser[2], Maarten Aerts[3], Peter Vandamme[3], Peter Dawyndt[1]

[1]Department of Applied Mathematics and Computer Science, Faculty of Sciences, Ghent University, Ghent, Belgium

[2]Laboratory for Protein Biochemistry and Biomolecular Engineering, Faculty of Sciences, Ghent University, Ghent, Belgium

[3]Laboratory for Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium;

Correspondences: Bart.Mesuere@UGent.be

Keywords: metaproteomics, tryptic peptides, biodiversity analysis, treemap visualization

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

Unipept (unipept.ugent.be) integrates a fast index of tryptic peptides built from UniProtKB (Wu *et al.* 2005) records with cleaned up information from the NCBI Taxonomy Database (Wheeler *et al.* 2004) to allow for biodiversity analysis of metaproteome samples. With Unipept, Users can submit tryptic peptides obtained from shotgun MS/MS experiments to which the application responds with a list of all UniProtKB records containing that peptide. The NCBI Taxonomy Database is used to compute the complete taxonomic lineage of every UniProtKB record in the result set. Subsequently, these lineages are combined to compute the common lineage of the submitted peptide. Of this common lineage, the most specific taxonomic node is determined as the lowest common ancestor (LCA) using a robust LCA scanning algorithm. The resulting information is visualized using an interactive JavaScript tree view that bundles all taxonomic lineages, accompanied with a comprehensible table that contains all matched UniProtKB records.

Users can also submit a list of tryptic peptides. In this case, the LCA is calculated for every submitted peptide as described above. These LCAs are then bundled into a frequency table and visualized on the results page using an interactive treemap [Figure 6]. This treemap displays hierarchical data in a multilayer histogram-like graphical representation. The squares in the treemap each correspond to a taxonomic node in the NCBI taxonomy, with their size proportional to the number of peptides having that taxonomic node as their LCA. The cleaned up hierarchy of the NCBI taxonomy is used to tile the squares according to their occurrence in the taxonomic lineages. These squares are color coded according to their taxonomic ranks. This graphical representation allows users to see at a glance which organisms are present in a metaproteome sample and to what extent. The treemap is interactive and can be manipulated by clicking on individual nodes. This makes it possible for users to zoom in to an area of interest (e.g. *Bacteria* or *Firmicutes*).

With complex samples containing a diverse range of taxa, the treemap representation quickly becomes cluttered. To resolve this problem, a new ring chart visualization was built into Unipept. Ring charts display the same data as the treemap, but as an interactive multi-level pie chart. The center of this pie chart represents the root node, with each ring around it stepping one level down the taxonomic hierarchy. The color of each slice is computed as the average of the colors of its children and slices without children are given random colors. Ring charts provide a more comprehensive view by displaying only four levels at a time. Users can see more levels by clicking on a slice of interest. The node that was clicked then

becomes the center of the ring chart and the four levels below it are displayed. By clicking on the center of a ring chart, users can zoom out one level. By hovering the mouse over a slice, a tooltip is displayed that gives more information about the taxonomic node associated with the slice. The tooltip shows the number of peptides that have the taxon as their LCA, and the number of peptides whose LCA is the taxon or one of its descendants in the NCBI taxonomy. These visualizations make Unipept an essential tool for gaining novel insights into the biodiversity of complex metaproteome samples [7-9].

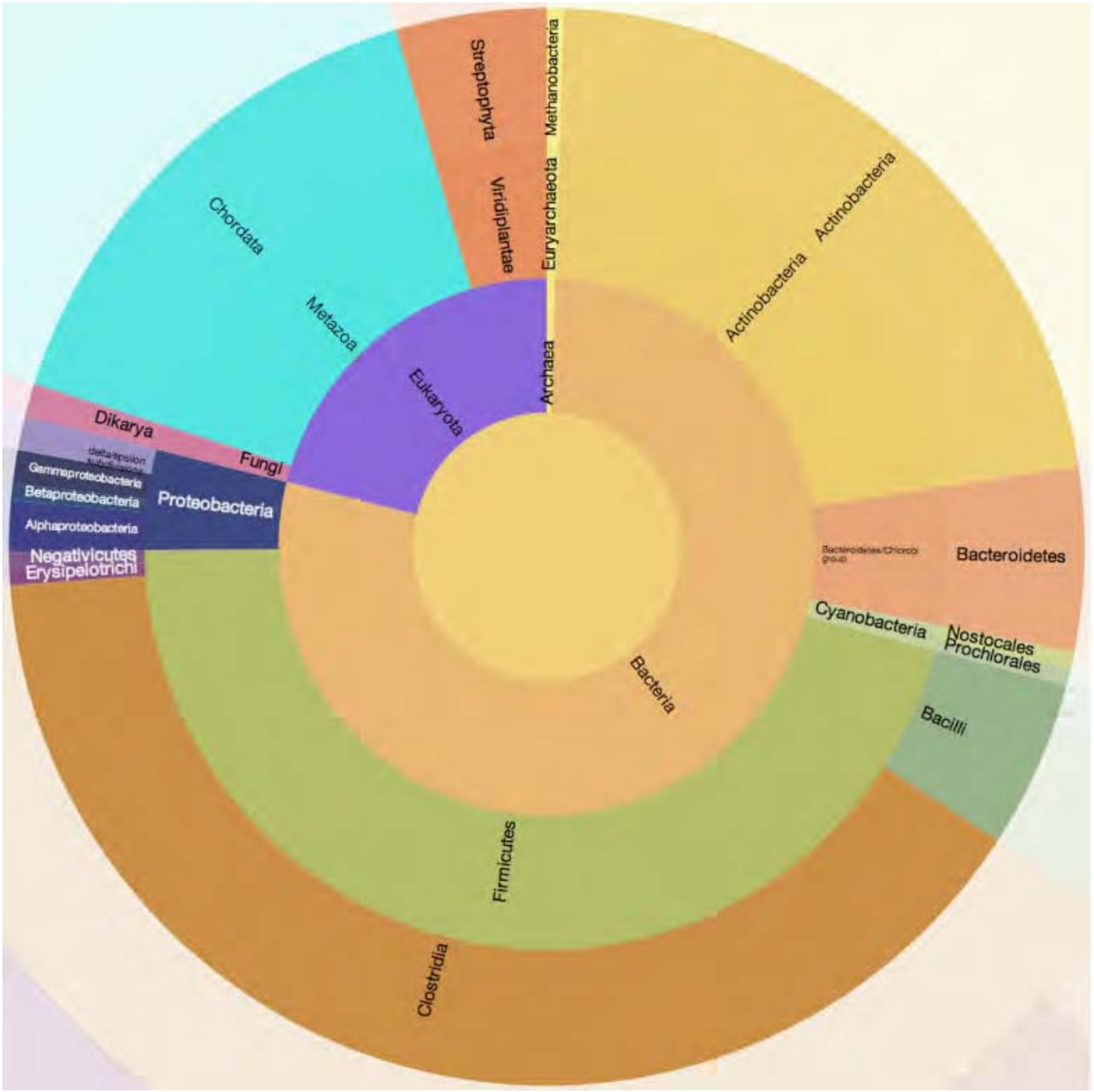


Figure 6. Ring chart visualization of treemap data.

Markus Göker

GSC15p09
Proposal for a Minimum Information on a Phenotype MicroArray Study (MIPS) standard

Markus Göker, Johannes Sikorski and Hans-Peter Klenk

Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

Correspondence: markus.goeker@dsmz.de

Keywords: minimum standards, phenotyping, genome annotation, gene function

Program session: Development of Resources, Tools or Databases Related to the GSC Mission, Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

The Phenotype MicroArray (OmniLog® PM) system developed by BIOLOG Inc. is able to simultaneously capture a large number of phenotypes by stimulation or inhibition of an organism's energy production over time with distinct substrates. The phenotypic reaction of single-celled organisms such as bacteria, fungi, and animal cells in nearly 2,000 assays can be tested in sets of 96-well microtiter plates to evaluate the response of cells to diverse environments. Processing these data includes parameter estimation from the respiration curves for quantitative analysis and discretization of these parameters into intrinsically positive or negative reactions for qualitative analysis. Phenotype MicroArray data are of use in genome annotation and the reconstruction of metabolic networks. Biochemical pathways inferred from genome annotations predict the abilities, or lack thereof, to metabolize certain compounds, or the resistance or susceptibility to certain substances. These hypotheses can be tested using the OmniLog® instrument. Genome annotation and pathway reconstruction can thus be iteratively improved. Projects such as the EU-funded MICROME currently investigate the application of OmniLog® measurements to genome-sequenced model organisms.

A minimum standard regarding the metadata recorded for, and distributed with, Phenotype MicroArray data has not yet been established. Because this kind of measurements does not represent genomic information but an increasingly important type of genome-associated data, it makes sense to establish such a minimum standard under the umbrella of the GSC. Another argument for a GSC project devoted to Phenotype MicroArray measurements is that the minimum standard for recording the organism-related metadata should be kept in sync with the organism-related part of MIGS. (A minority of OmniLog® users apply it to phenotyping environmental samples; minimum standards for such experiments should be kept in sync with MIMARKS and MIMS.)

We have implemented and published OPM, a package for the free statistical software environment R that offers tools for storing the curve kinetics, aggregating the curve parameters, recording associated metadata of organisms and experimental settings as well as methods for analyzing these highly complex data sets graphically and statistically. It is also possible to discretize and export these parameters. Export and import in a standardized YAML format already facilitates the data exchange among labs. It would be easy to include automated checking of MIPS compliance in this package, and it can serve as a software exemplar for applying the novel checklist. Once MIPS is established, it would soon be possible to distribute metadata-enriched Phenotype MicroArray datasets in standardized file formats.

Peter Dawyndt

GSC15p10
ALFALFA: a Novel Algorithm for Long Fragment Mapping and Alignment

Michael Vyverman, Bernard De Baets, Veerle Fack and Peter Dawyndt

Ghent University, Ghent, Belgium

Keywords: read mapping, sequence alignment, long reads, sparse suffix arrays

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

Mapping and alignment of sequencing reads against reference sequences is a crucial step in many genome analysis pipelines. Due to the continuous growth in sequencing data, the development of fast and accurate read mapping algorithms with a limited memory footprint has received much attention in recent years. However, most algorithms focus on short reads and are not well equipped for the longer reads produced by a growing number of sequencing platforms. With the advent of third generation sequencing featuring ever longer reads, there is definitely a need for faster, memory-friendlier and more accurate long read mapping tools.

We developed ALFALFA, a new Algorithm for Long Fragment Alignment. ALFALFA presents a novel combination of tested algorithmic techniques and heuristical filtering techniques. It is based on the classical seed-and-extend approach and utilizes sparse enhanced suffix arrays to compute its seeds. Sparse suffix arrays have similar time-memory trade-offs as their widely used compressed counterparts, but are much faster when memory is abundant and also benefit from improved cache effects when memory is scarce.

ALFALFA has been compared against other long read mappers, including Bowtie2 and BWA-SW on the independent Rabema benchmark, using both real and simulated data. The results show practical speed-memory-accuracy trade-offs for all mapping algorithms. However, for most experimental set-ups, ALFALFA is four to five times faster than other long read mappers, while maintaining equally low memory footprint and slightly higher accuracy.

Anjette Johnston

GSC15p11
GSC-Compliance at NCBI

Anjanette Johnston, John Anderson, Tanya Barrett and Ilene Mizrachi

NCBI, National Institutes of Health, Bethesda, Maryland, USA

Keywords: BioProject, BioSample, metadata, GSC-compliance, NCBI

Session: Development of Resources, Tools or Databases Related to the GSC Mission, Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

Submitting contextual metadata along with sequence data to NCBI's Primary Data Archives is important for providing users with a complete understanding of the source of the biological data. We worked with the GSC to promote the inclusion of descriptive metadata within sequence submissions to GenBank in a tabular structured comment. GSC-compliant metadata was validated in submission tools and flagged with a GSC-specific keyword. This information is now captured in two databases designed specifically to house contextual metadata, BioProject for project-specific information, and BioSample for sample-specific attributes.

BioProject is a resource that aggregates linkages to various components of a research project. Collections of projects exist that include links to retrieve and browse information related to a specific initiative. These links can be based on funding source, overall goal, organism type. etc., and can support multiple relationships between related or

diverse BioProjects. All submitted data are linked by a common BioProject ID, which allows for retrieval in Entrez and through a tabular summary. BioProject also includes links to BLAST resources, SRA data, and BioSample.

BioSample is a central location in which to store normalized, descriptive information about biological source materials used to generate experimental data. BioSample provides a single point of submission for metadata that may be referenced when making data deposits to archival databases. In order to promote collection of biologically useful information, BioSample defines "attribute packages" that drive provision of specific information appropriate to the sample type. Currently, General and Pathogen packages are supported as well as packages compliant with GSC MIxS standards. Controlled vocabularies are promoted for sample attributes, thus helping to harmonize sample descriptions across NCBI.

BioSample currently serves as the master record for source metadata provided for SRA and BioProject and, eventually, will function as the central repository of source information for most of NCBI's Primary Data Archives. This will allow users to aggregate all available data present in multiple archival databases that are derived from a sample with common attributes, as well as view them in the context of their associated BioProjects.

Evangelos Pafilis

GSC15p12
ENVIRONMENTS: identification of environment descriptive terms in text

Evangelos Pafilis[1]*, Sune Frankild[2], Lucia Fanini[1], Sarah Faulwetter[1], Christina Pavloudi[1], Katerina Vasileiadou[1], Christos Arvanitidis[1], Lars Juhl Jensen[2*]

[1]Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research, Crete, Greece

[2]Disease Systems Biology, Novo Nordisk Foundation for Protein, University of Copenhagen, Copenhagen, Denmark

*Correspondence to: Evangelos Pafilis (pafilis@hcmr.gr), Lars Juhl Jensen (lars.juhl.jensen@cpr.ku.dk)

Abstract

ENVIRONMENTS is an open source, dictionary-based, named entity recognition tool supporting such term identification.

Summary

Community Standards promote data exchange and facilitate large-scale integrative biological analysis. While structured information available in biological databases can be used to this end, the discovery potential increases once input based on the knowledge existing in the scientific literature is considered.

The identification of environment descriptive terms, such as "terrestrial", "aquatic", "lagoon", "coral reef", in text is a prerequisite for mining crucial environmental context information.

ENVIRONMENTS is an open source, dictionary-based, named entity recognition tool supporting such term identification. The Environment Ontology's (EnvO, http://environmentontology.org) controlled and structured vocabulary for biomes, environmental features, and environmental materials, serves as the source of names and synonyms for such identification process.

Built on the software infrastructure of tools characterized, among others, by fast performance (Pafilis *et al.* [10]) ENVIRONMENTS is capable of addressing the challenge posed by the ever increasing biomedical, ecology and biodiversity literature.

Orthographic dictionary expansion and flexible matching are improving the matches between EnvO terms as they exist in the ontology and the way they may be written in text. An extensively manually curated stopword list is safe-guarding against increased false positives.

Applying ENVIRONMENTS to biomedical, ecology and biodiversity related literature can be employed to characterize biological entities such as genetic sequences (on-going work in collaboration with Dr. C. Quince, Dr. U. Ijaz et al. in the context of http://www.cost.eu/domains_actions/essem/Actions/ES1103) and species.

EP and LF have received funding from the European Union's Seventh Framework Programme

(FP7/2007-2013) under grant agreement No 264089 (MARBIGEN project). LJJ and SuF by the Novo Nordisk Foundation Center for Protein Research. A visit of EP in NNFCPR has been funded by an EMBO Short Term Fellowship (356-2011).

Peter Sterk

GSC15p13
The Metagenomics Resource at the European Bioinformatics Institute

Hubert Denise, Peter Sterk, Matt Corbett, Matthew Fraser, Craig McAnulla, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Maxim Scheremetjew and Sarah Hunter

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

Keywords: EBI, SRA, MGPortal, GSC, standards, metagenomics, analysis, InterPro

Program Session: Development of Resources, Tools or Databases Related to the GSC Mission, Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

The EBI Metagenomics resource is a single access point to submit, analyze, visualize and retrieve metagenomics and metatranscriptomics datasets. EBI Metagenomics allows users to search and access public metagenomic or metatranscriptomic samples by accession number or via plain text search. They can then visualize and download analysis results or download the raw data from the Sequence Read Archive (SRA) on the European Nucleotide Archive (ENA) website.

Experimentalists can upload their own data for archiving in the Sequence Read Archive (SRA) using a web-based data submission system – SRA Webin. Archiving data in this way is frequently a requirement for journal publication. The submission and archiving process requires a simple registration step, and allows users to keep their data confidential prior to publication.

Metadata are important to accurately describe the samples and the experiments performed. In order to encourage the submission of this essential contextual data, study-type-specific checklists that comply with Genomic Standards Consortium (GSC; http://gensc.org) guidelines have been implemented in the SRA Webin submission tool.

After submission of the data to SRA, the EBI Metagenomics resource generates a functional profile for each sample using InterPro to analyze the predicted protein coding sequences. The approach here is to use computational models (protein signatures) in databases such as Pfam, PROSITE, PRINTS, CATH-Gene3D and TIGRFAMs to identify protein families, domains and functionally important sites within protein sequence fragments. We are introducing a taxonomic diversity analysis of both amplicon and whole genome shotgun data using adaptations of pipelines from the Qiime package (http://qiime.org).

In the near future we hope to improve the searchability of standards-compliant metadata to facilitate reuse and reinterpretation of metagenomic data.

[No photo]

GSC15p14
GSC 15 abstract

Susanne Schneider

Michigan State University, East Lansing, MI, USA

Correspondence: George Garrity (garrity@msu.edu)

A MIxS and DwC compliant biodiversity Reference Data Model to support the community in implementing the database layer

Abstract

The main goal is to bring macroscale, microscale, and physical, morphological and sequence-based biodiversity observation together under one concept, by building a MIxS/DwC and Open Geospatial Consortium standards compliant Reference Data Model for storing geo-referenced genome and metagenome sequences and associated contextual metadata, as well as traditional biodiversity observational data, which presents the harmonization of the genomic and biodiversity standards.

Based on this Reference Data Model, one or more reference implementations will be created including physical data models for Oracle and PostgreSQL. This allows the community members easily to create consistent, interoperable local applications for their own purposes. Moreover, given the size, heterogeneity, and complexity of sequence and biodiversity data, the Reference Data Model will allow a precise discussion among developers about the best implementation possible for a given database management system. The Reference Data Model also serves as basis for data transport, and ways of exchanging data. A RESTful Web API specification will be developed for web services, which will focus on specifying HTTP methods and resource oriented URLs for data requests and corresponding responses.

The responses will be defined for different data formats and ensure consistency and full coverage of the DwC and MIxS standards. Furthermore, tools for data import/export will be developed, particularly focused on adaptors for data transformations between the reference implementations and common data exchange formats, such as Genomic Contextual Data Markup Language (GCDML), Comma Separated Value (CSV), and Resource Description Framework (RDF). With the GCDML already in hand, the reference implementation would use GCDML first as the data format for the API response definition. However, to fulfill users' needs for getting data in different formats, we will develop both logical descriptions and technical tools to support several automated transformations from GCDML to other formats. A first priority is RDF, which is particularly important for integrating with DwC. Also simple formats like CSV require support, given the ubiquity with which they are used in practice.

To this end, we will create application-agnostic specifications and documentations with examples of use for developers of specific applications, which will enhance understanding and communication of the standardized infrastructure and ensure consistency of standards-compliant data. It will also minimize the chance of multiple cycles of re-approximating needs and, while minimizing conflicting versions, allow better reuse of community tools. In addition, data sharing will be easier and could be accomplished in a loss-free manner, with higher performance and lower costs, because the need for complex model transformations would be minimized. Finally, it will allow users, such as molecular ecologists, to work more readily with sequences greatly enriched with standards-compliant environmental data from different sources. Ideally, the investigators would be able to do so without having to know about the underpinning standardized infrastructure.

Taken together, the implementation would lead to a significantly higher degree of interoperability of data and service providers on all layers, thereby also raising common understanding, providing an educational framework, and ensuring more efficient informatics development.

Finally, we aim to bring this project under the umbrella of the Genomic Standards Consortium and hence invite the community to participate.

Susanna-Assunta Sansone

GSC15p15
The ISA Commons: curate, visualize, analyze, share and publish

Philippe Rocca-Serra, Eamonn Maguire, Alejandra Gonzalez-Beltran and Susanna-Assunta Sansone

University of Oxford e-Research Centre, Oxford, UK;

Correspondence: isatools@googlegroups.com

Abstract

We describe the work of a global network of like-minded communities, using the Investigation / Study / Assay (ISA) open source metadata tracking framework to facilitate standards-compliant *collection*, *curation*, *visualization*, *management*, *sharing, publication* and *reuse* of datasets in an increasingly diverse set of life science domains, including metabolomics, (meta)genomics, proteomics, system biology, environmental health, environmental genomics and stem cell discovery. We also illustrate our next steps to improve management of bioscience data in the cloud; use of semantic web approaches to make existing knowledge available for linking, querying, and reasoning; link to existing open source analysis tools; new model in scientific publishing. The vision of the ISA framework is to support the gradu-

al progression from unstructured, usually non-digital, experimental descriptions to structured information that - when shared - is comprehensible, reproducible and reusable.

Rationale

A growing worldwide movement for reproducible research encourages making data, along with the experimental details available in a standardized manner. Also several data management, sharing policies and plans have emerged in response to increased funding for high-throughput approaches in genomics and functional genomics science [11]. In parallel, a growing number of community-based groups are developing hundreds of standards (minimal reporting requirements, terminologies and exchange formats) to harmonize the reporting of different experiments, so that these can be comprehensible and, in principle, reproduced, compared and integrated [12]. But data annotation is a time-consuming task. User-friendly, open source tools are needed to (i) empower researchers or curators (supporting them) to use relevant community standards when collecting and describing the experimental information, (ii) store, query, integrate the experiments and submit them to public repositories, where relevant, (iii) support reasoning on and analysis of the data, and also (iv) publish it.

A global network of like-minded communities

At the heart of the ISA Commons [13] there is the general-purpose ISA-Tab file format [14], built on the 'Investigation' (the project context), 'Study' (a unit of research) and 'Assay' (analytical measurement) metadata categories. The extensible, hierarchical structure of this format enables the representation of studies employing one or a combination of assays and technologies, focusing on the description of its experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships), using relevant community-standards. The ISA software suite [14,15]- the second element of this framework – acts to create and edit ISA-Tab files, store, serve and convert them to a growing number of related formats.

The ISA Commons' community is delivering a growing ecosystem of public and internal resources (that use the ISA-Tab file format, and/or is powered by one or more component of the ISA software suite) ranging from *international public repositories* [16], *institutional repositories* [17] to *funded research consortia* [18] and *data journals* [19,20]. These variety of implementations, illustrate the flexibility of the format and the versatility of the ISA framework that belongs to its community of users and contributors, assisted by a dedicated team (authors of this article) who has led and supported its open developments since 2007.

K. Baker

GSC15p16
The Global Genome Biodiversity Network (GGBN) and the Global Genome Initiative (GGI): Building the Infrastructure for the future of Genomics

K. Barker[1], S. Brady[1];J. Coddington[1], G. Droege[2], C. Meyer[1]

[1]Smithsonian Institution National Museum of Natural History, Washington, DC, USA

[2]Botanic Garden and Botanical Museum Berlin-Dahlem, Berlin, Germany

Abstract

Biodiversity science is increasingly focused on studying taxonomic groups at the genomic level, as shown by the Genome10k, I5k, and 1Kite projects. Access to genomic specimens becomes especially important, as progress in DNA-based technologies continues to speed up research on the genetic diversity of life forms. Nevertheless, the scientific community still lacks routine, standard knowledge of, and access to, publically available collections of genome quality material, comparable to phenotypic collections as stored in natural history museums over the last several hundred years. The Global Genome Initiative aims to fill the gaps in existing collections through collaborative collecting that respects access and benefit sharing (ABS), and to foster the creation of public biorepositories that publish their holdings through one virtual, global, biorepository portal (the Global Genome Biodiversity Network). Crucial to the latter is one international metadata

standard for genomic tissue and/or DNA collections. Here we present progress on the development of GGI field protocols for collecting genome quality tissues across the Tree of Life as an output of the GGI 2013 Capturing Genomes Workshop and progress on the development of the GGBN data portal and the standards for providing access to genome quality samples and related data (e.g. underlying voucher specimen, GenBank entries, etc), as outputs of the GGBN 2012 Meeting and TDWG 2012 Conference, respectively.

Cynthia Parr

GSC15p17
Encyclopedia of Life: Applying concepts from Amazon.com and LEGO to biodiversity informatics

Cynthia Parr

Smithsonian Institution, Washington DC, USA

Keywords: biodiversity, ecology, evolution, data standards, semantics

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

After more than five years building its initial infrastructure for aggregating descriptive information on all living biodiversity, the Encyclopedia of Life (EOL, http://www.eol.org) is undergoing a major transformation in order to better serve scientific discovery. We will present preliminary results from three linked data pilot studies, we will outline a variety of new studies that will use various standard vocabularies and methods for mining EOL for structured data, and we will introduce EOL's methods for harvesting and managing structured data, such as trait data, for biodiversity scientists. These efforts are important for increased interoperability of biodiversity and genomic datasets.

Several pilot studies are underway using EOL taxon identifiers and other methods of linking data across domains. Pete Devries and Anne Thessen are mining EOL text data for species names in order to extract potential species associations. The data will be semantically annotated using DBPedia Spotlight, and thus connected to the rest of the Linked Open Data cloud. Rod Page and Ryan Schenk are building BioNames, which links EOL pages, literature from Biodiversity Heritage Library and elsewhere, and phylogenies from PhyLoTA. Finally, a group of zooarchaeologists is linking fossil specimens using EOL identifiers so that they can trace the history of animal domestication in the near East.

Approximately seven newly-funded projects under the Rubenstein program will use a variety of API calls, text mining, and crowdsourcing methods to extract, link, and visualize a variety of numeric and controlled vocabulary values from EOL pages in order to answer large-scale questions in ecology, evolution, and conservation biology.

Together with a Sloan-funded effort to support marine biodiversity use cases, these projects are all helping to refine requirements for how Encyclopedia of Life can manage and serve to users useful data for scientific analyses. We are building a generalized, flexible, scalable system that weds the current taxonomic and coarse-grained subject infrastructure (based largely on Darwin Core and the TDWG Species Profile Model) to the more detailed data standards used by other databases. The idea is to foster high-level integration and querying along with preservation of low-level semantics. One could search for all data relevant to physical description, for all data using a particular EnvO term, or for all data available for a collection of taxa whether on EOL or in an external database like GenBank or GBIF. EOL will perform taxonomic name resolution and serve the requested data along with specific term identifiers (e.g. URIs), source attributions, and curation status. Third-party developers will be encouraged to develop tools for extracting new data, for integrating with other databases, and for finding pattern and meaning in the overall data sets.

As hundreds of providers to the Encyclopedia of Life converge on standards for describing species and linking these species to information in related resources, we will enable large-scale discovery and retrieval for cross-domain projects. The general approach can be compared to the Amazon.com and LEGO models, where a general standard enables a thriving marketplace connecting data consumers with a wide variety of possible providers who have building blocks for biodiversity studies.

G. Sutton

GSC15p18
Standards for Pan-Genome Data Repositories

G. Sutton

J. Craig Venter Institute

Keywords: pan-genome, standards, databases

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

The availability of cheap sequencing has enabled the sequencing and comparison of many strains of the same species of bacteria. The study of the gene repertoire and differences in gene content between these strains has been called pan-genome analysis. A key step in this analysis is identifying the operationally equivalent genes across these strains. This is a necessary step to determine differences in gene content. A multiple sequence alignment of operationally equivalent genes enables analysis of potentially functionally significant polymorphisms. Anchoring an alignment of the genomes at the operationally equivalent genes provides information on genome rearrangements and differences in upstream and downstream regions of the genes that could contribute to changes in regulation. Genome alignments also inform what regions are prone to the gain or loss of regions of DNA and which regions tend to stay intact perhaps to maintain tight regulatory control of an important biological system. The layout of the genes can be just as important as the gene content. Public repositories will need to represent pan-genome information as well as the underlying genome information. What amount of pan-genome information to store, how that information is arrived at, and standards for computing, storing, and displaying that data need to be determined. Two obvious benefits for the public repositories from pan-genome clustering are: data reduction and consistency of annotation. We try to elucidate the outstanding issues for pan-genome data repositories and propose some standards to deal with them.

H.H. Creasy

GSC15p19
The Human Microbiome Project Metadata Catalog

HH Creasy

Institute for Genome Sciences, Baltimore, MD, USA

Keywords: human microbiome, metagenomics, genomic metadata

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

The NIH Human Microbiome Project Data Analysis and Coordination Center (HMP DACC) provides an infrastructure to enable the scientific community to access, analyze and interpret human microbiome data, with the ultimate goal of advancing our understanding of human health. The HMP DACC website, available at hmpdacc.org, provides background details on various aspects of the project, as well as terabytes of HMP reference genome, metagenomic 16S and wgs data. The DACC works closely with the Genomes Online Database (GOLD) and the GSC to provide MIxS-compliant genomic and metagenomic metadata via the HMP Project Catalog. Efforts are underway to release compliant metadata through NCBI BioSample. This is being accomplished through data download and manual retrieval, parsing, and manual curation. Here we will discuss our achievements in curating HMP metadata, as well as informing users as to how they can attain all of this critically important data.

Nikos Kyrpides

GSC15p20
The Genomes OnLine Database (GOLD) v.5: status of genomic and metagenomic projects and their associated metadata.

K. Liolios and N. Kyrpides

DOE, Joint Genome Institute, Walnut Creek, CA, USA

Keywords: GOLD, metadata, standards

Abstract

The Genomes OnLine Database (GOLD, http://www.genomesonline.org) is a comprehensive resource for centralized monitoring of genome and metagenome projects worldwide. Both complete and ongoing projects, along with their associated metadata, can be accessed in GOLD through precomputed tables and a search page. As of December 2012, GOLD, now on version 5.0, contains information for 19,527 sequencing projects, of which 4050 have been completed and their sequence data has been deposited in a public repository. Out of these complete projects, 2,365 are finished and 1,685 are permanent drafts. Moreover, GOLD contains information for 358 metagenome studies associated with 2,231 metagenome samples. GOLD continues to expand, moving toward the goal of providing the most comprehensive repository of metadata information related to the projects and their organisms/environments in accordance with the Minimum Information about any (x) Sequence specification and beyond.

K. Eric Wommack

GSC15p21
The VIROME Compare-inator: a web tool for exploring the diversity of viral communities

K. Eric Wommack, Dan Nasko, Jaysheel Bhavsar and Shawn Polson

University of Delaware, Baltimore, MD, USA

Keywords: shotgun metagenomics, microbial ecology, viral ecology

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

Cataloging and censusing populations within communities of organisms is fundamental to many ecological investigations. For cellular microorganisms, polymorphism within the universally-shared small subunit rRNA gene has become a central component of ecological studies. For viruses no universal marker gene exists; thus, researchers have turned to shotgun metagenomic approaches as a means of cataloguing and censusing viral populations. We have developed the Viral Informatics Resource for Metagenome Exploration (VIROME), a bioinformatics pipeline and web-application dedicated to the analysis and visualization of shotgun viral metagenome sequence data. In the VIROME pipeline, genetic features within metagenome libraries are characterized through a comprehensive BLAST analysis against the UniRef100 (UR100) database and MetaGenomes On-Line (MGOL), a custom annotated database of environmental peptides. Hits to UR100 peptides are cross-referenced with six other annotated databases to provide detailed characterization of the frequency of gene functions and taxa within a viral community. Through hits to MGOL peptides, viral metagenome libraries are characterized according to the environmental distribution of BLAST homologs. These results provide researchers with a rich collection of annotations and metadata for building matrices of functional, taxonomic, and environmental observations that characterize autochthonous viral communities within environmental samples. We have developed a tool called the

Compare-inator that builds observation matrices from BLAST data against functional databases (SEED, KEGG, COG, ACLAME, GO), the NCBI taxonomy database, and MGOL environmental annotations to enable quantitative comparative analyses of viral communities. The Compare-inator exports frequency observation data in both tab-delimited and the biological observation matrix format (BIOM), as well as a comprehensive metadata file for libraries in the matrix. The BIOM and metadata file is suitable for import into the Quantitative Insights in Microbial Ecology (QIIME) analysis package, providing a straightforward path for comparative analysis of viral communities. In this presentation we demonstrate use of the Compare-inator for time series analysis of dynamic changes in viral communities.

Patricia L. Whetzel

GSC15p22
NCBO Technology- Powering Semantically Aware Applications to enable Standards Driven Research

Patricia L. Whetzel and NCBO team

Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA

Abstract

Researchers have turned to ontologies and Semantic Web technologies to annotate and integrate disparate knowledge. The use of ontologies in standard reporting guidelines is critical to provide a common language for researchers to describe biomedical investigations and to drive these processes. The successful creation of standards-based semantic applications in the life sciences requires Web services that provide access to ontologies. The National Center for Biomedical Ontology (NCBO), a National Center for Biomedical Computing created under the NIH Roadmap, developed BioPortal, which provides access to the largest repository of biomedical ontologies available online (http://bioportal.bioontology.org). The ontology content in BioPortal is programmatically accessible through a standard interface via the NCBO's Web services. These Web services enable the development of semantically aware applications for data annotation, data integration, and natural language processing. The NCBO Web services can be grouped into four categories: *Ontology Access*, *Mapping*, *Annotation*, and *Data Access*. The *Ontology Access* Web services provide access to ontologies, their metadata, ontology versions, downloads, navigation of the class hierarchy (parents, children, siblings), and details of each term. The *Mapping* Web services provide access to the millions of ontology mappings published in BioPortal. The *NCBO Annotator* Web service "tags" text automatically with terms from ontologies in BioPortal, and the *NCBO Resource Index* Web services provides access to an ontology-based index of public, online data resources. The NCBO Web services have been incorporated into over 50 applications thus far, including those developed by members of the Genomics Standard Consortium such as the BioSamples database, the BioSharing project, ISAcreator, MG-RAST, OntoMaton, and QIIME. This presentation will describe the NCBO Web services and applications using these services driving the development of a semantic infrastructure within the genomics community.

Daniel Nasko

GSC15p23
Assessing patterns of CRISPR spacer composition using the CASC discovery and validation tool

Daniel Nasko[1], Jaysheel Bhavsar[2], Shawn Polson[1] and K. Eric Wommack[1]

[1]University of Delaware, Newark, DE, USA

[2]Mayo Clinic, Phoenix, AZ, USA

Keywords: CRISPRs, Bioinformatics, Metagenomics, Viral Ecology

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

Although increasingly detailed information on the mechanistic basis of CRIPSR immunity has been uncovered, the impact of CRISPRs on natural viral

communities remains largely unknown. In particular, little is known about the identity of phage genes that contribute to spacer sequences. Fortunately the characteristic repeats of the CRISPR locus make it possible for pattern matching algorithms to predict putative novel spacers. While several CRISPR discovery tools already exist (e.g. CRISPR Finder, CRT, PilerCR) they all share a common pitfall of high false positive rates. In an effort to mitigate these false positive rates, we have created CASC (CASC Ain't Simply CRT) a program for prediction and validation of CRISPR spacers in genomic and metagenomic sequences. Using the Viral Informatics Resource fOr Metagenome Exploration (VIROME) web application, we classified viral metagenomic open-reading frames (ORFs) showing homology to spacer sequences. Two independent spacer datasets were used: known spacers from the CRISPR-finder database (ca. 60,000 spacers) largely derived from whole genomes of cultivated microbes; and a smaller collection of novel, hypothetical spacers identified by CASC in microbial metagenomes, including those from the Global Ocean Sampling (GOS) (ca. 3,000 spacers). In homology searches against the publicly available VIROME viral metagenomic libraries, the novel spacers from microbial metagenomes were ~ 100-fold more likely to show homology to a viral metagenomic ORF than known genomic spacers from the CRISPR-finder database. Spacer targets from both datasets showed high frequencies of ORFs with an assignable function, as well as ORFs with mixed viral/bacterial homology (expected as spacers are present in both), differing significantly from the distribution of VIROME ORF classes within the component shotgun viral metagenomes. Additionally, BLAST alignments show evidence that some select sets of uncharacterized viral ORFs contain highly conserved regions, which appear to be preferentially targeted to become spacers. Together these data tend to indicate that a smaller and perhaps more select group of viral genes are likely to become retained targets within CRISPR immune system arrays.

Hilmar Lapp

GSC15p24
The blessing and the curse: handshaking between general and specialist data repositories

Hilmar Lapp and Todd Vision

National Evolutionary Synthesis Center (NESCent), Durham, NC, USA

University of North Carolina, Chapel Hill, NC, USA

Keywords: data sharing ecosystem, metadata standards, metadata exchange, repository interoperability

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

The growing number of both general and specialist data repositories and the advent of a multitude of discipline-specific metadata standards hold out the promise of a diverse interoperable ecosystem of reusable research data curated to community standards. At the same time, they pose the risk of a fragmented ecosystem that is confusing to users and that does not fully reward the investment of the research community. For researchers, increasingly often some combination of repositories exists for all the kinds of data they are expected to archive from a study; however, identifying the appropriate combination of general and specialist repositories can be bewildering, and the burden of provisioning (sometimes redundant) metadata to specialist repositories can hinder uptake. General repositories are challenged to incentivize provision of adequate discipline-specific metadata, while non-mandated specialist repositories are challenged to incentivize deposit at all. Both kinds of repositories are challenged to maintain linkages to related research products hosted externally (e.g. literature, software, and other data).

We consider ways to enjoy the blessings of this complex ecosystem while warding off its curses. In metagenomics, almost all journals mandate sequence data deposition in an INSDC database. In some cases, these databases also take sample and study metadata and short-read sequences. Specialist metagenomics repositories with tailored querying, analysis, and visualization features exist, too, but operate more or less independently.

Metagenomics studies often produce other long-tail data that may be archived in a general repository such as DataDryad. We propose mechanisms, and invite feedback, on how a general repository could promote: (1) minimum reporting standards for sequencing and metagenomics experiments, (2) the uptake of not only mandated but also community-standard specialist repositories, (3) the archiving of long-tail data that falls in the gaps between specialist repositories and (4) efficient metadata exchange among all the repositories hosting data for a given study. Our proposals are informed by lessons learned from an experiment in handshaking between DataDryad and the TreeBASE phylogenetics database.

Takamoto Fujisawa

GSC15p25
Development of an ontology for the INSDC Feature Table Definition

Takatomo Fujisawa[1], Toshiaki Katayama[2], Shinobu Okamoto[2], Hiroshi Mori[3], Yasunori Yamamoto[2] and Yasukazu Nakamur[1]

[1]National Institute of Genetics, Research Organization of Information and Systems, Mishima, Japan

[2]Database Center for Life Science, Research Organization of Information and Systems, Tokyo, Japan

[3]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Tokyo, Japan

Keywords: semantic web, RDF, genome ontology, INSDC, DDBJ

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

Semantic Web technology makes it possible to improve data integration and retrieval more efficient. Since 2006, the Semantic Web technology is introduced in the UniProt protein knowledge base and has been used in the operation and management of the database. In Japan, we are promoting the integration of databases using Semantic Web technology in DDBJ, DBCLS and NBDC.

During the NBDC/DBCLS international BioHackathon 2012 held in Japan, developers of life science databases and applications discussed and agreed on to develop new ontology for describing locations of the objects on a sequence. As a result, the Feature Annotation Location Description Ontology (FALDO) was proposed and we converted the locations and positions of all annotations stored in our triple store to comply with this new standard. Also, in domestic BioHackathon BH12.12, we have built the working-draft for an ontology of the DDBJ/EMBL/GenBank Feature Table Definition, which is the common annotation document revised by the INSDC once a year. We aim to build an ontology for standardized, systematic description of the INSDC sequence entry which includes the information of submitters, references, source organisms, and the biological feature annotations.

[No Photo]

GSC15p26

Sequencing and analysis standards for microbial forensics

Adam M. Phillippy

National Bioforensic Analysis Center, National Biodefense Analysis and Countermeasures Center, Frederick, MD, USA

Correspondence: phillippya@nbacc.net

Abstract

Microbial forensics entails the characterization and comparison of microbial samples in support of criminal investigations. The 2001 anthrax letter attacks pushed the field towards whole-genome sequencing as the preferred method for microbial typing due to its unparalleled resolution and accuracy in analyzing DNA polymorphisms. In addition, high-throughput sequencing has enabled new metagenomic approaches to forensics and pathogen detection. There is substantial overlap between these forensic methods and those utilized by the microbial epidemiology, ecology, and evolution research communities; however, forensics demands a much higher degree of certainty and standardization. This has necessitated the validation and standardization of routine genomics approaches such as whole-genome sequencing and metagenomics.

Towards this goal, the National Bioforensic Analysis Center (NBFAC) recently achieved ISO 17025 accreditation for whole-genome sequencing and bioinformatics analysis, and is leading a number of initiatives to advance the capabilities of genomics-based microbial forensics. In particular, I will review our ISO program and recent efforts to develop and validate software for the assembly of genomic and metagenomic sequencing data, including the GAGE [1] and MetAMOS [2] projects. Future plans include the extension of these efforts to include whole-genome SNP typing and metagenomic classification methods. I will also outline future challenges and potential areas for collaboration between the forensics community and international organizations such as the Genomic Standards Consortium.

Renzo Kottman

GSC15p27
Earth Sampling App for Ocean Sampling Day

Renzo Kottmann[1], Julia Schnetzer[1,2], Aleksandar Pop Ristov[3], Frank Oliver Glöckner[1,2]

[1]Max Planck Institute for Marine Microbiology, Microbial Genomics and Bioinformatics Research Group, Bremen, Germany

[2]Jacobs University Bremen, School of Engineering and Sciences, Campusring 1, 28759 Bremen, Germany

[3] InterWorks, Bitola, Republic of Macedonia

Abstract

More than two thirds of the contextual data required by the Minimum Information about any Sequence (MIxS) standard can already be obtained while sampling on site. The Earth Sampling App is a mobile application which records sampling data at time of collection in the field. This application is specifically tailored for the needs of the Ocean Sampling Day. The aim is to support scientists and citizens in optimal gathering of GSC compliance data, already at the beginning of a complex experimental workflow. Moreover, to comply with the discussions and suggestions of the "Field Information Management Systems Hackathon" at GSC 14, the Earth Sampling App development may also support the standardization of simultaneous field data transport to multiple receiving servers.

Folker Meyer

GSC15p28
Pan-genome: a new paradigm in microbiology

Folker Meyer[1], Tatiana Tatusova[2] and David Ussery[3]

[1]Argonne National Laboratory, Argonne, IL, USA

[2]National Center for Biotechnology Information, U.S. National Library of Medicine 8600 Rockville Pike, Bethesda MD, 20894 USA

[3]Technical University of Denmark

Keywords: microbial genome, pan-genome, comparative genomics

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

After the first complete genome of a free-living organism, *Haemophilus influenzae*, was sequenced in 1995 [PMID 7542800], whole-genome sequencing became a standard method for the study of the biological processes in living forms. At the beginning of the genomic era, it was thought that a single representative isolate was sufficient to describe the genetic complexity of a species, and the use of 'comparative genomics' was restricted to investigating the diversity among different yet closely related bacteria. More recently, multiple isolates of the same species have been sequenced and analyzed. It is now known that intra species variation can be as significant as interspecies diversity. Bacterial genomes from various strains of the same species can vary considerably in genome size, nucleotide composition and gene content. It has become clear that, bacterial species cannot be

represented by an individual reference strain or a set of reference genomes. The 'pan-genome' concept has been introduced by Tettelin et al in 2005 [reference PMID 16172379]. The pPan-genome has been defined as a super-set of all genes in all the strains of a species. A pPan-genome includes the "core genes" that are present in nearly all strains, "accessory genes" present in two or more strains, and finally "unique genes" specific to single strains. A shift in the paradigm from individual genome to 'pan-genome' has occurred in the past few years, with the rapid advances in the sequencing technology. The main approach in pan-genome studies is a comparative analysis of multiple strains from a single species, although one can also describe pan-genomes for different taxonomy levels – for example, a phylum or genus pan-genome, or sometimes even a subspecies pan-genome (as in the case of E. coli O157:H7, with 34 genomes sequenced so far). Alternative approaches include the use of rapidly growing meta-genomic sequence data and single-cell genome sequencing. The pan-genome concept is already changing the way we understand bacterial evolution, adaptation, and population structure, and has further important implications in identification of virulence genes. But the data model for a pan-genome concept is yet to be determined and the standards for the data and metadata data formats are yet to be defined.

Judith A. Blake

GSC15p29
Gene Ontology: Functional anotation fo cmparative gnomics

Judith A.Blake and Suzanna Lewis

The Jackson Laboratory, Bar Harbor, ME, USA

The Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Abstract

The Gene Ontology (GO) provides the community standard for knowledge representation of functional information about proteins in both cellular and viral systems. The domain specific ontologies provided by the GO Consortium are the most widely used structured controlled vocabulary for representing and accessing knowledge about proteins. Since the first use of the Gene Ontology in 1999 for functional annotations of the newly sequenced genomes of the major model organisms yeast (*Saccharomyces cerevisiae*), fruitfly (*Drosophila melanogaster*) and laboratory mouse (*Mus musculus*), the GO system has been used to provide authoritative information from biomedical literature about the functioning of over 350,000 proteins from a wide-range of species. Through comparative genomics approaches, these core annotations permit controlled inference of functional information for nearly 350,000 species and over 96 million gene products. The GO curation effort continues to develop and mature both in ability to mine biomedical literature for key annotations for proteins under experimental investigation, and in the ability to provide high-quality and controlled annotations for all other proteins.

Key to the successes of the GO effort are the contributions from the global community of developers and curators of genome databases and repositories as well as from bioinformaticians and research groups that contribute their domain expertise to ensure the completeness and quality of the knowledge representations. The GO Consortium brings together ontology developers and annotation teams virtually and physically at regular and ad hoc meetings to improve the ontology structures and to expand curation in selected areas such as apoptosis or cardiac conduction.

Recently improvements in GO include incorporation of additional relations and structures to increase expressivity including when and where proteins are active. We have developed a phylogenetic approach to propagating experimentally-based annotations via comparative analysis. Going forward, we are working with the model organisms database community to provide a common annotation tool.

We will present an overview of GO program and progress. We will discuss how the GO community has developed a system that provides standards for knowledge representation that have been adopted and incorporated in major genomics and bioinformatics resources.

Junhua Li

GSC15p30
A new version of human gut microbial gene catalog integrating with 3 continent populations and human gut-related bacterial genomes.

Junhua Li[1], Shinichi Sunagawa[2], Xianghang Cai[1] and Huanzi Zhong[1]

[1]BGI Research, Szhenjeng, China

[2]EMBL, Heidelberg, Germany

Keywords: human gut microbiome, gene catalog, metagenomic sequencing, reference dataset, update and improvement

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

With the development of culture-dependent and culture-independent technology, much more human gut microbiome communities from distinct populations and isolated bacterial genomes have been exposed. Recent studies showed that the diversity of human gut microbiota is beyond our previous estimation, which means that we still need more investigation among wide scale population(Huttenhower et al., 2012a; Lozupone, Stombaugh, Gordon, Jansson, & Knight, 2012; Yatsunenko et al., 2012). Additionally, human gut microbe related disease association study or other comparison study require an integrated and as complete as possible gene catalog as reference. Hence, we extended the human gut microbial gene catalog published by MetaHIT project(Qin & Li, 2010) from 124 to 760 Europeans. Further, we gathered stool sample metagenomic sequencing data from two recently published human gut microbiome studies, which are 139 American samples from HMP (Methé et al., 2012)and 368 Chinese samples from a Chinese T2D study(Qin et al., 2012). To capture the human gut inhabit but rare (low abundance) bacteria as much as possible, genes from sequenced bacterial genome were useful and necessary to be included. In order to accurately define human gut-related bacteria, we took three aspects of bacteria into consideration, and selected 511 human gut-related bacteria genomes from about 3500 published bacteria genomes data set.

In general, we processed totally about 6.4Tb metagenomic sequencing data from 1,267 individuals with standardized pipeline MOCAT(Kultima et al., 2012) to establish gene catalog for each population, and integrated them with genes from 511 human gut-related bacteria. Finally, an integrated and the most comprehensive by far human gut microbiome reference gene catalog was established, containing nearly 10 million prevalent genes with average length of 753bp.

To access the improvement of our new integrated gene catalog constructed by larger scale samples, distinct populations, we compared it with previous published catalogs(Qin & Li, 2010) on gene sequences, functional pathway integrity and species integrity level. Rarefaction curve analysis showed that saturation is very close, and the ratio of completeness estimated by Chao2 and ICE was 94.5% and 95.35% respectively. Comparison on the gene content, the integrated gene catalog covered 87.8% gene content of previous one with 24.3% genes of itself, which meant 75.6% genes in new catalog are novel. At the advantage of plenty novel genes found, the amount of reads which could be represented by catalog was improved by 10% on average. Functional comparison results suggested that the pathway integrity was modestly improved, and it is consist with previous study showed the stability of functional composition of gut microbiome(Huttenhower et al., 2012b). However, obvious improvement on species genome coverage was noticed even though we only considered the gene catalog from 3 metagenomic sequencing cohorts (3CGC), which was without complement of sequenced bacterial genome. In terms of 987 species/strains genome of which were covered by more than 10% by any catalog, 723 of them were increased by 10.0%-80.4%, 35.0% on average, especially in *Enterococcus*, *Bifidobacterium*, *Lactobacillus*, *Streptococcus* and *Enterobacteriaceae*.

To summarize, the integrated gene catalog, compared with previous published one, has achieved outstanding improvement, and could sever as a much better reference for following human gut microbiome researches.

Elizabeth M. Glass

GSC15p31
Using Data Standards in KBase: The DOE Systems Biology Knowledgebase

Adam Arkin[1], Robert Cottingham[1], Sergei Maslov[3], Rick Stevens[4], Thomas Brettin[4], Dylan Chivian[1], Paramvir Dehal[1], Christopher Henr[4], Folker Meyer[4], Jennifer Salazar[4], Doreen Ware[5], David Weston[2], Brian Davison[2] and Elizabeth M. Glass[4]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[2]Oak Ridge National Laboratory, Oak Ridge, TN, USA

[3]Brookhaven National Laboratory, Upton, NY, USA

[4]Argonne National Laboratory, Argonne, IL, USA

[5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

Keywords: Systems biology, Knowledgebase, Data integration, Data exchange

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

KBase is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It will be a community-driven, extensible and scalable open-source software framework and application system. Our immediate 18-month goal is to have a beta-version completed by February 2013.

The KBase microbial science domain will enable the reconciliation of metabolic models with experimental data with the ultimate aim of manipulating microbial function for applications in energy production and remediation The plants science domain will initially target linking genetic variation, phenotypes, molecular profiles, and molecular networks, enabling model-driven phenotype predictions. We will also map plant variability onto metabolic models to create model-driven predictions of phenotypic traits. Our microbial communities team will build the computational infrastructure to research community behavior and build predictive models of community roles in the carbon cycle, other biogeochemical cycles, bioremediation, energy production, and the discovery of useful enzymes.

KBase will offer free and open access to data models and simulations, enabling scientists and researchers to build new knowledge, test hypotheses, design experiments, and share their findings to accelerate the use of predictive biology. In order to integrate and exchange diverse biological data types, KBase will use data standards provided by various organizations and consortiums like that of the GSC.

Lynn Schriml

GSC15p32
Meta Data Standards for the Built Environment - MIxS-BE

Elizabeth M. Glass[1], Yekaterina Dribinsky[2], Pelin Yilmaz[3], Hal Levin[4], Robert Van Pelt[5], Doug Wendel[6], Andreas Wilke[1], Jonathan Eisen[7], Sue Huse[8], Anna Shipanova[9], Mitch Sogin[8], Jason Stajich[9], Rob Knight[10], Folker Meyer[1] and Lynn Schriml[6]

[1]Argonne National Laboratory, Argonne, IL, USA

[2]Max Planck Institute for Marine Microbiology, Bremen, Germany

[3]Building Ecology Research Group, Santa Cruz, CA, USA

[4]Private Practicing Architect, Boulder, CO, USA

[5]University of Colorado, Boulder, CO, USA

[6]University of California Davis, Davis, California, USA

[7]Marine Biological Laboratory, Woods Hole, MA, USA

[8]University of California, Riverside, Riverside, CA, USA

[9]University of Maryland School of Medicine, MD, USA

Keywords: Built Environment, Meta data package, Microbial communities, Indoor

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

The composition of indoor microbial communities has the potential to profoundly affect human health. A number of factors within a building or room can alter the microbial abundance and diversity, such as

occupancy, temperature, and humidity, which in turn impacts the quality of the air we breathe. The Sloan Foundation has established the Microbiology of the Built Environment (BE) to uncover the complexity of microbial ecosystems of inside spaces. Bringing together researchers and architects, we have established a working group to expand the GSC MIxS standard for microbial sequences collected from Built Environments. Samples collected, sequenced and annotated with MIxS-BE metatdata from waste-water, air filters, air and surfaces of indoor spaces provides a rigorous and structured tool for analysis of microbial sequences and ecosystems of the indoor and outdoor environments. These packages include a minimal list and a more comprehensive list of terms. The BE metadata package has been integrated into such resources as MG-RAST and QIIME for use by the metagenomics community. This provides a venue for researchers to use the BE package and provide feedback. We will continue to work with the community to evolve this standard to describe room properties and properties unique to specific types of buildings, to better serve their needs. This is especially important as scientists discover more factors that impact the microbial communities of the built environment.

Pelin Yilmaz

GSC15p33
Phylogeny, ecology and biogeography of marine bacterial and archaeal clades

Pelin Yilmaz, Renzo Kottmann and Frank Oliver Gloeckner

Max Planck Institute for Marine Microbiology, Bremen, Germany

Keywords: MIxS, metadata, marine, metagenome

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

In this study, we investigated the phylogeny and taxonomy of mostly uncultivated (meaning no or very few cultured relatives) marine bacterial and archaeal clades on a global scale, in the form of a meta-analysis. Known marine bacterial and archaeal clades (e.g. SAR11, SAR86, Gammaproteobacterial OMG clades), which are already annotated on the SILVA SSU rRNA reference tree (Quast et al. 2012), were connected to their original publications, and these sequences were subjected to phylogenetic reconstruction methods in order to determine their exact phylogenetic/taxonomic affiliations. Additionally, these "published" marine clades, which are limited to 20-30 groups, were extended by text mining methods to determine groups that are of solely (or mostly) of marine origin, again using the SILVA rRNA databases.

In the second part of this work, we used these marine clades sequence dataset in a large scale meta-analysis of all available metagenome and amplicon (16S variable regions). The datasets chosen for this analysis were based on MixS metadata supplied. The results of this analysis were used in an effort to gain insights into the global distribution of various marine clades, their ecology, biogeography, and interaction with oceanographic variables, with an ultimate aim of creating a "field guide" of our knowledge of uncultured marine bacterial and archaeal clades.

George M. Garrity

GSC15p34
Standards In Genomic Sciences – An Open Access Journal of the Genomic Standards Consortium

George M Garrity and Oranmiyan Nelson

Michigan State University, East Lansing, MI, USA

Keywords: Open Access Publishing, Genome Sequence Announcement, Genomic Standards, Core Project

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

Standards in Genomics Sciences (SIGS; www.standardsingenomics.org) was founded as an experimental open access publication to promote the data standardization efforts of the Genomic Standards Consortium (GSC). Whereas peer-reviewed publication of genome sequences was commonplace at the outset of the genomics era, many of the established journals in the life sciences abandoned such papers as the number of sequenced genomes increased sharply, leading to a loss of contextual information needed to correctly analyze and interpret genome sequence data. SIGS uses highly structured, easy to read and edit reports of genome and metagenome sequences, standard operating procedures, meeting reports, white papers and other articles that are relevant to a growing readership. Highlights for 2012 included SIGS receiving its first Impact Factor (1.63) and effective publication of a number of new species and genus descriptions that incorporate genome sequence data into the descriptions. At the time of writing, SIGS had published 280 articles, including 234 short genome reports. Our readership continues to grow, topping 64,000 readers in 171 countries, representing a 71% increase in audience in 2012 during 2012. SIGS is listed in CrossRef, PubMedCentral, Scopus, the Web of Science, ChemAbstracts, EBSCO and the Directory of Open Access Journals (DOAJ) and has become one of top three journals publishing papers on new genome sequences.

Ilene Mizrachi

GSC15p35
Standards in the INSDC

Ilene Mizrachi
National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda MD, USA

Abstract

The INSDC is committed to archiving and distributing high quality sequences and annotation along with rich metadata to describe the project and biological sample. BioProject and BioSample databases have been developed to facilitate the capture of structured metadata for diverse biological research projects and samples represented in NCBI's archival databases. NCBI and its INSDC partners promote the use of standards for describing samples, experimental methodology and annotation. The adoption of common standards by the databases and scientific community will simplify the exchange of knowledge between different resources and improve the utility of the data for making scientific discoveries. Through a newly developed Submission Portal (http://submit.ncbi.nlm.nih.gov) for the deposition of experimental data and associated metadata, NCBI can enforce rules to ensure that annotation and metadata standards are being met for as a condition of submission.

Petra ten Hoopen

GSC15p36
ENA tools for sample metadata compliance

Petra ten Hoopen, Blaise Alako, Clara Amid, Lawrence Bower, Ana Cerdeño-Tárraga, Iain Cleland, Richard Gibson, Neil Goodgame, Mikyung Jang, Simon Kay, Rasko Leinonen, Xiu Lin, Arnaud Oisel, Nima Pakseresht, Swapna Pallreddy, Sheila Plaister, Rajesh Radhakrishnan, Stéphane Rivière, Marc Rossello, Alexander Senf, Nicole Silvester, Dmitriy Smirnov, Ana Toribio, Daniel Vaughan, Vadim Zalunin and Guy Cochrane

European Nucleotide Archive, European Molecular Biology Laboratory-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

Keywords: metadata, standards compliance, ENA, INSDC, MIxS, MicroB3

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

Harmonization of (meta)data collection becomes an essential effort in the age when data generation is often easier and more affordable then

their organization and storage. Minimum sample information that should accompany archived nucleotide sequences was first specified by the INSDC (International Nucleotide Sequence Database Collaboration) and then further extended by MIGS/MIMS/MIMARKS standards. These checklists, collectively named MIxS, were developed by the GSC (http://gensc.org/gc_wiki/index.php/Main_Page) as mechanisms to standardize description of (meta)genomes and marker genes. The three standards share common core descriptors, differ in checklist-specific elements and can be tailored to a particular environment by a subset of relevant environment-specific information components.

New tools recently developed at the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) extensively support harmonized reporting of sequenced-sample metadata. Assembled and annotated sequences and their metadata can be submitted and validated via submission-type-specific templates. Samples of sequence reads are described and captured information validated via 16 checklists comprising of the INSDC default checklist, the complete list of GSC MIxS terms and all its environmental subsets, the ENA prokaryotic pathogen checklist and the MicroB3-Sample checklist. This last checklist supports standards being developed by the MicroB3 Project Consortium and shall assure a sample metadata compliance not only with the MIxS standards but also with minimal reporting requirements for oceanographic data, advancing the metadata compliance and their interoperability across different research domains. The MicroB3-Sample checklist will facilitate description of samples collected during the Ocean Sampling Day, a simultaneous sampling campaign of the world's oceans to reveal a marine microbial diversity.

Our work on user-friendly submission technology and workflows is driven by our belief that such tools can have a profound impact on standards compliance and data discoverability and utility.

Jack Gilbert

GSC15p37
The Earth Microbiome Project and the importance of data standards in modeling the planet's microbiome

Jack Gilbert[1], Janet Jansson[2] and Rob Knight[3]

[1]Argonne National Laboratory, Argonne, IL, USA

[2]Lawrence Berkley National Laboratory, Berkeley, CA, USA

[3]University of Colorado at Boulder, Boulder, CO, USA

Keywords: Microbiome, Microbiology, Metagenomics, EMP, Earth, Modeling

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

The understanding of Earth's climate and ecology requires multi-scale observations of the biosphere, of which microbial life are a major component. However, to acquire and process physical samples of soil, water and air that comprise the appropriate spatial and temporal resolution to capture the immense variation in microbial dynamics, would require a herculean effort and immense financial resources dwarfing even the most ambitious projects to date. To overcome this hurdle we created the Earth Microbiome Project, a crowd-sourced effort to acquire physical samples from researchers around the world that are, importantly, contextualized with physical, chemical and biological data detailing the environmental properties of that sample in the location and time it was acquired. The EMP leverages these existing efforts to target a systematic analysis of microbial taxonomic and functional dynamics across a vast array of environmental parameter gradients. The EMP uses the MIxS data standards format to capture the environmental gradients, location, time and sampling protocol information about every sample donated by our valued collaborators. Physical samples are then processed using a standardized DNA extraction, PCR, and shotgun sequencing

protocol to generate comparable data regarding the microbial community structure and function in each sample. To date we have processed >10,000 samples, and have >20,000 in the process of being analyzed. One of the key goals of the EMP is to map the spatiotemporal variability of microbial communities to capture the changes in important processes that need to be appropriately expressed in models to provide reliable forecasts of ecosystem phenotype across our changing planet. This is essential if we are to develop economically sound strategies to be good stewards of our Earth. The EMP recognizes that environments are comprised of complex sets of interdependent parameters and that the development of useful predictive computational models of both terrestrial and atmospheric systems requires recognition and accommodation of sources of uncertainty.

Rob Knight

GSC15p38
The Hospital Microbiome Project – understanding microbial transmission and ecology in a health care environment

Jack Gilbert[1], Daniel Smith[1], Brent Stephens[2], Jeffrey Siegel[3], Hal Levin[4], Rob Knight[5], Emily Landon[6], Stephen Weber[6], Sylvia Garcia-Houchins[6], Michael Morowitz[7], John Alverdy[6], Jessica Green[8], Scott Kelley[9] and Benjamin Kirkup[10]

[1]Argonne National Laboratory, Argonne, IL, USA

[2]Illinois Institute for Technology, Chicago, IL, USA

[3]University of Toronto, Toronto, Ontario, Canada

[4]Building Ecology Group, Santa Cruz, CA, USA

[5]University of Colorado at Boulder, Boulder, CO, USA

[6]University of Chicago Hospital, IL, USA

[7]Childrens' Hospital Pittsburgh, PA, USA

[8]University of Oregon, Eugene, OR, USA

[9]San Diego State University, San Diego, CA, USA

[10]US Army

Keywords: Hospital, Microbiome, Metagenomics, Sequencing, Modeling

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

Despite their role in healthcare provision, very little is known about how microbial (e.g. bacteria, Archaea and fungi) organisms are transmitted around the built infrastructure of a hospital (this knowledge limitation also applies to other built environments, but is especially critical in a hospital context). The Hospital Microbiome Project is characterizing the taxonomic composition of surface-, air-, water-, and human-associated microbial communities in two hospitals to monitor changes in community structure following the introduction of patients and hospital staff, or major surgical procedures. The goal is to determine the influence of numerous factors on the rate and nature of microbial community succession in these hospitals including: human population demographics, how these demographics interface with a space, and the building materials, environmental conditions, and building operational characteristics used to create and maintain that space. This ongoing initiative is taking place in a newly constructed private US hospital in Chicago and a US Army medical center in Germany. In Chicago, 12,392 samples are being collected using sterile swabs from patients, staff, rooms, common areas, water, and air filters from 52 or 365 time points prior to and following the official opening of the hospital. In Germany, 1600 samples are being collected from surfaces in one military room over 16 time points, with 100 samples collected per time point. Absolute microbial abundance (plate counts and qPCR) and building environmental measurements (ventilation rates, temperature, relative humidity, light intensity, and human occupancy) are being combined with relative taxonomic and functional gene abundance via amplicon sequencing (16S/18S/ITS) and shotgun metagenomics. Here we will present preliminary data from this initiative and describe the role of data standards, especially the Minimal Information about ANY sequence (MIxS) Built Environment descriptions and the use of the BIOM format for data sharing and exchange, in enabling this massively collaborative initiative.

Susan Holmes

GSC15p39
Reproducible research for data normalization, analyses and simulations

Susan Holmes and Paul J. McMurdie

Stanford University, Stanford, CA, USA

Keywords: Reproducible Research, Simulations, Bootstrap, Confirmatory Analyses

Program session: Development of Resources, Tools or Databases Related to the GSC Mission, Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

Collaborative Platforms providing a way for researchers to compare different protocols for normalizing and analyzing Microbiome data using R will be presented. We will also highlight some of the difficulties inherent in making choices between different types of distances and normalization techniques and how to carry out reproducible robustness and sensitivity analyses.

We provide examples of dealing with data heterogeneity, whether from different technologies or different data types.

Curtis Huttenhower

GSC15p40
Cataloging genes and pathways in the human microbiome

Nicola Segata[1], James Kaminski[2] and Curtis Huttenhower[2]

[1]University of Trento

[2]Harvard School of Public Health

Keywords: microbiome, metagenome, metatranscriptomes, orthologous groups,pathways

Program session: Implementing Standards in Genomic and Metagenomic Research Projects

Abstract

Several major studies of the human microbiome, including the HMP and MetaHIT, have produced catalogs of microbial gene families accumulated from hundreds of metagenomes. Nearly every additional metagenomic and metatranscriptomic study also relies on the definition of gene or protein families of interest and their identification from meta'omic reads. Many orthologous family catalogs appropriate for this task exist, as well as many methods for providing systematic identifiers to family members encountered in new meta'omes. Finally, these gene families are collected both into unstructured biological process groups and into structured pathways by a diversity of functional catalogs. I will discuss our experience with gene family identification and pathway reconstruction during and after the HMP and suggest standardizations necessary for these tasks in the future.

Steven L. Salzberg

GSC15p41
The Challenges of Big Data in Next-Generation Genomics

Steven L. Salzberg

Director, Center for Computational Biology

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Correspondence:Steven L. Salzberg (http://bioinformatics.igm.jhu.edu/salzberg)

Abstract

Next-generation sequencing (NGS) technology allows us to peer inside the cell in exquisite detail, revealing new insights into biology, evolution, and

disease that would have been impossible to find just a few years ago.

Summary

The enormous volumes of data produced by NGS experiments present many computational challenges that we are working to address. In this talk, I will discuss solutions to two basic alignment problems [21]: mapping sequences onto the human genome at very high speed, and [22] mapping and assembling transcripts from RNA-seq experiments. I will also discuss some of the problems that can arise during alignment and how these can lead to mistaken conclusions about genetic variants, RNA editing, and changes in gene expression.

My group has developed algorithms to solve each of these problems, including the widely-used Bowtie and Bowtie2 programs for fast alignment [21,22] and the TopHat and Cufflinks programs for assembly and quantification of genes in transcriptome sequencing (RNA-seq) experiments [23,24]. This talk describes joint work with current and former group members including Ben Langmead, Cole Trapnell, Daehwan Kim, and Geo Pertea; and with collaborators including Mihai Pop and Lior Pachter.

Bing Ma

GSC15p42
Organism-centric approach to study paired metagenome and metatranscriptome profiles of vaginal microbial community

Bing Ma[1], Arthur Brady[1], Owen R. White[1], Jacques Ravel[1]

[1]The Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA

Abstract

Metatranscriptome analyses have mostly relied on mapping sequencing reads to reference genome databases. Because of the lack of robust evaluation tools, highly variable sequence coverage, and demanding computational power, *de novo* metatranscriptome assembly are challenging.

In this study, we sequenced on the Illumina HiSeq platform paired metagenomes and metatranscriptomes from human associated microbiota to explore employing metagenome information for metatranscriptomic analyses by first generating species-level metagenome assemblies to map metatranscriptome sequence reads. We optimized the performance of *de bruijn* graph-based metagenome assemblers to generate assemblies with high fidelity and low chimericity. The k-mer abundance distribution was estimated in digital normalization: 1) reads where sequence coverage was over 50× (redundant information) and less than 5× (most likely erroneous reads) were removed; 2) reads were then partitioned based on a *de Bruijn* graph connectivity to optimize local coverage prior to assembly. Contigs were binned by species, such that when the majority of the reads making a contig mapped to a bacterial species, the contig was assigned to that species.

We were able to obtain high quality assemblies from strains of bacteria species with 75-95% sequence similarity to reference genomes. Further, nearly complete genome sequences can be constructed with as little as 3% of total reads (~180,000 reads). Our method allows for the recruitment into contigs of an additional ~5-50% reads that could not be mapped to any reference genomes. When using these high quality reconstructed genome sequences as template to map metatranscriptomic sequence data as much as 25% more sequence reads were recruited for any given bacterial species when compared to the number of sequence reads map directly reference genomes. We applied our methods to dissect species level metatranscriptomics functional activities of vaginal bacterial species.

Owen White

GSC15p43
Utilization of Standardized Vocabularies in the Management of Human Microbiome Data

Owen White, Lynn Schriml, Michelle Gwinn, Heather Hout, Jonathan Crabtree, Victor Felix and Anup Mahurkar

Institute for Genome Sciences, University of Maryland, Baltimore, MD, USA

Keywords: metagenomics, data warehouse, metadata

Program session: Development of Resources, Tools or Databases Related to the GSC Mission

Abstract

The Human Microbiome Project Data Analysis and Coordination Center has been responsible for the management of microbiome samples derived from >700 healthy individuals as well as >15 pilot projects examining several disease states (http://hmpdacc.org). During the course of these projects we made considerable use of the Minimum Information about: -Marker Gene Sequence (MIMARKS), -Metagenomic Sequence (MIMS) and -Genome Sequences (MIGS) specifications to maintain metadata for 16S rRNA Whole Metagenome Shotgun, and reference strain bacteria respectively. Our experience with using these systems, as well as metadata to the relevant human phenotypic information associated with the microbiome projects will be discussed. In addition to this effort, we have developed the Open Science Data Framework (OSDF, osdf.igs.umaryland.edu) a scalable data warehousing system. We have deployed an instance of OSDF to manage nearly all of the HMP data. OSDF uses schemas that are implemented using JSON, which are closely linked to both standardized metadata (e.g., MIGS) and custom metadata formats. We will discuss the potential for our current implementation of these schemas to support an exchangeable internationally standardized schema file format.

References

1. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* 2011; **29**:415-420. PubMed http://dx.doi.org/10.1038/nbt.1823

2. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 2011; **39**:W339-W346. PubMed http://dx.doi.org/10.1093/nar/gkr466

3. Medema MH, Breitling R, Bovenberg R, Takano E. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nat Rev Microbiol* 2011; **9**:131-137. PubMed http://dx.doi.org/10.1038/nrmicro2478

4. Medema MH, van Raaphorst R, Takano E, Breitling R. Computational tools for the synthetic design of biochemical pathways. *Nat Rev Microbiol* 2012; **10**:191-202. PubMed http://dx.doi.org/10.1038/nrmicro2717

5. Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. and Glöckner, F.O. (2012) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res, online, Quast, C., Pruesse, E.,

6. Yilmaz A, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* (In press). PubMed

7. Mesuere B, Devreese B, Debyser G, Aerts M, Vandamme P, Dawyndt P. Unipept: tryptic Peptide-based biodiversity analysis of

metaproteome samples. *J Proteome Res* 2012; **11**:5773-5780. PubMed

8. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, *et al*. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* 2004; **32**:D35-D40. PubMed http://dx.doi.org/10.1093/nar/gkh073

9. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, *et al*. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 2006; **34**:D187-D191. PubMed http://dx.doi.org/10.1093/nar/gkj161

10. Pafilis E, Frankild SP, Fanini L, Faulwetter S, Pavloudi C, Vasileiadou K, Arvanitidis C, Jensen LJ. The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS ONE* (In press).

11. Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, *et al*. 'Omics data sharing. *Science* 2009; **326**:234-236. PubMed http://dx.doi.org/10.1126/science.1180598

12. Sansone SA, Rocca-Serra P. On the evolving portfolio of community-standards and data sharing policies: turning challenges into new opportunities. *GigaScience* (2012).

13. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, *et al*. Toward interoperable bioscience data. *Nat Genet* 2012; **44**:121-126. PubMed http://dx.doi.org/10.1038/ng.1054

14. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, *et al*. ISA software suite. *Bioinformatics* 2010; **26**:2354-2356. PubMed http://dx.doi.org/10.1093/bioinformatics/btq415

15. Maguire E, González-Beltrán A, Whetzel PL, Sansone SA, Rocca-Serra P. OntoMaton: bringing semantic annotation to Google spreadsheets for collaborative data management. *Bioinformatics* 2013; **29**:525-527. PubMed http://dx.doi.org/10.1093/bioinformatics/bts718

16. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, *et al*. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* 2013; **41**:D781-D786. PubMed http://dx.doi.org/10.1093/nar/gks1004

17. Ho Sui SJ, Begley K, Reilly D, Chapman B, McGovern R, Rocca-Sera P, Maguire E, Altschuler GM, Hansen TA, Sompallae R, *et al*. The Harvard Stem Cell Discovery Engine. *Nucleic Acids Res* 2012; **40**:D984-D991. PubMed http://dx.doi.org/10.1093/nar/gkr1051

18. ToxBank, pan-European consortium. http://toxbank.net/about/datawarehouse

19. BMC GigaScience journal. gigadb.org/mouse-methylomes

20. New Nature Publishing Group data product - to be launched early 2013.

21. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**:R25. PubMed http://dx.doi.org/10.1186/gb-2009-10-3-r25

22. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**:357-359. PubMed http://dx.doi.org/10.1038/nmeth.1923

23. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; **25**:1105-1111. PubMed http://dx.doi.org/10.1093/bioinformatics/btp120

24. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010; **28**:511-515. PubMed http://dx.doi.org/10.1038/nbt.1621

Meeting participants

Full Name (First Middle Last)	Job Title	Companyorganization	Email
Omayma Al-Awar	Manger, Sequencing Specialist	Illumina	oalawar@illumina.com
Marc Allard	microbiologist	FDA CFSAN ORS	marc.allard@fda.hhs.gov
Linda Amaral Zettler	Associate Scientist	Marine Biological Laboratory	amaral@mbl.edu
John Anderson	Staff Scientist	NIH/NLM/NCBI	anderson@ncbi.nlm.nih.gov
Katharine Barker	GGI Coordinator	Smithsonian Institution	barkerk@si.edu
Tanya Barrett	Staff Scientist	NCBI, NLM, NIH	barrett@ncbi.nlm.nih.gov
Reed Beaman	Assoc. Curator	University of Florida	rbeaman@gmail.com
Mesude Bicak	Ocean Sampling Day Coordinator / Bioinformatician	University of Oxford	mesude@gmail.com
Matthew Bietz	Assistant Project Scientist	University of California, Irvine	mbietz@uci.edu
Judith Blake	Associate Professor	The Jackson Laboratory	judith.blake@jax.org
Jason Bobe	Executive Director	PersonalGenomes.org	jasonbobe@gmail.com
Vivien Bonazzi	Program Director	NHGRI	bonazziv@mail.nih.gov
Jing Chen	Programmer Analyst	UCSD	jic002@ucsd.edu
Jonathan Coddington	Associate Director for Science	National Museum of Natural History, Smithsonian Institution	coddington@si.edu
James R Cole	Assistant Professor	Michigan State University	colej@msu.edu
Rita Colwell	Distinguished University Professor	University of Maryland	rcolwell@umiacs.umd.edu
Heather Creasy	Sr Bioinformatics Analyst	Institute for Genome Sciences	hhuot@som.umaryland.edu
Lawrence David	Junior Fellow	Harvard University	ldavid@fas.harvard.edu
Neil Davies	Director	UC Berkeley - Moorea	ndavies@moorea.berkeley.edu
Peter Dawyndt	Prof. Dr.	Ghent University	Peter.Dawyndt@ugent.be
John Deck	Programmer	UC Berkeley	jdeck@berkeley.edu
Daniel Drell	Program Manager	Biological and Environmental Research, DOE	daniel.drell@science.doe.gov
Robert Dunn	Associate Professor	Department of Biology, North Carolina State University	rroberdeaudunn@gmail.com
Scott C Edmunds	Editor	GigaScience/BGI	scott@gigasciencejournal.com
Dawn Field	Professor	NERC Centre for Ecology and Hydrology	fiedawn@googlemail.com
Roberto Flores	Program Director	National Cancer Inst.	floresr2@mail.nih.gov
Takatomo FUJISAWA	researcher	National Institute of Genetics	tf@nig.ac.jp
George Garrity	Professor, Editor-in-Chief	Michigan State University	garrity@msu.edu

Meeting participants (cont.)

Full Name (First Middle Last)	Job Title	CompanyOrganization	Email
Jack Gilbert	Environmental Micorbiologist	Argonne National Laboratory	gilbertjack@uchicago.edu
Frank Oliver Glökner	Groupleader	Max Planck Institute for Marine Microbiology	fog@mpi-bremen.de
Elizabeth Glass	Bioinformatics Engineer	Argonne National Laboratory	marland@mcs.anl.gov
Markus Goeker	Dr.	DSMZ	markus.goeker@dsmz.de
Antonio Gonzalez Pena	Research Associate	University of Colorado at Boulder	antgonza@gmail.com
Susan Gregurick	Program Manager	DOE	susan.gregurick@science.doe.gov
Lynette Hirschman	Director, Biomedical Informatics	The MITRE Corporation	lynette@imtre.org
Susan Holmes	Professor	Stanford	susan@stat.stanford.edu
Chris Hunter	Lead Biocurator	GigaScience	only1chunts@gmail.com
Curtis Huttenhower	Assistant Professor	Harvard School of Public Health	chuttenh@hsph.harvard.edu
Grace Hwang	Principal Biosensors Scientist	MITRE Corporation	gmhwang@mitre.org
Anjanette Johnston	Staff Scientist	NIH/NLM/NCBI	johnston@ncbi.nlm.nih.gov
William Klimke	Staff Scientist	NCBI	klimke@ncbi.nlm.nih.gov
Anna Klindworth	Postdoc	Max Planck Institute for Marine Microbiology	aklindwo@mpi-bremen.de
Rob Knight	Associate Professor	HHMI / University of Colorado	rob.knight@colorado.edu
Tonia Korves	Multi-Discipline Sys Eng, Sr	The MITRE Corporation	tkorves@mitre.org
Renzo Kottmann	Post-Doc	MPI for Marine Microbiology	rkottman@mpi-bremen.de
Nikos Kyrpides	Senior Staff Scientist	DOE Joint Genome Institute	nckyrpides@lbl.gov
Yemin Lan	Research Assistant	Drexel University	yeminlan@gmail.com
Hilmar Lapp	Asst. Director for Informatics	National Evolutionary Synthesis Center (NESCent)	hlapp@nescent.org
Katja Lehmann	Bioinformatician	Centre for Ecology and Hydrology	kleh@ceh.ac.uk
Junhua LI	Group Leader	BGI	lijunhua@genomics.cn
Ann Lichens-Park	National Program Leader	U.S. Department of Agriculture	apark@nifa.usda.gov
Konstantinos Liolios	Group Lead	DOE Joint Genome Institute	kliolios@lbl.gov
Bing Ma	Postdoc	Institue for Genome Science at University of Maryland	bma@som.umaryland.edu
Marnix Medema	PhD Student	University of Groningen	m.h.medema@rug.nl
Christopher Meyer	Research Zoologist	Smithsonian Institution	meyerc@si.edu

Meeting participants (cont.)

Full Name (First Middle Last)	Job Title	CompanyOrganization	Email
Folker Meyer	Computational Biologist	Argonne National Laboratory	folker@mcs.anl.gov
Ilene Mizrachi	GenBank Coordinator/Primary Data Archives Section Chief	NCBI/NLM/NIH	mizrachi@ncbi.nlm.nih.gov
Hiroshi Mori	Assistant Professor	Tokyo Institute of Technology	hmori@bio.titech.ac.jp
Norman Morrison	Senior Data Scientist	The University of Manchester	norman.morrison1@gmail.com
Jayne Morrow	Research Engineer	NIST	jmorrow@nist.gov
Daniel Nasko	Graduate Research Assistant	University of Delaware	dnasko@udel.edu
Jim Ostell	Branch Chief	Information Engineering Branch, NCBI, NLM, NIH	ostell@ncbi.nlm.nih.gov
Andrea Ottesen	Research Microbiologist	FDA	andrea.ottesen@fda.hhs.gov
Evangelos Pafilis	Research Fellow	HCMR, Hellenic Center Marine Research	pafilis@hmcr.gr
Cynthia Parr	EOL Chief Scientist	Smithsonian Institution	parrc@si.edu
Adam Phillippy	Principal Investigator	NBACC	phillippya@nbacc.net
Shawn Polson	Assistant Professor	University of Delaware	polson@dbi.udel.edu
Lita Proctor	Program Director	NHGRI/NIH	lita.proctor@nih.gov
Patricia Reichelderfer	Extramural Program Staff	NIH/NICHD	reichelp@exchange.nih.gov
Robert Robbins	Scientist	UC San Diego	rjr8222@gmail.com
Gail Rosen	Asst. Professor	Drexel University	empr3ss@gmail.com
Marc Salit	Group Leader	NIST	marc.salit@nist.gov
Steven Salzberg	Professor	Johns Hopkins University	salzberg@jhu.edu
Susanna-Assunta Sansone	PI, Team Leader, Data Consultant	University of Oxford - and - Nature Publishing Group	sa.sansone@gmail.com
Michael Schneider	Ph.D. Student	Max Planck Institute for Marine Microbiology	mschneid@mpi-bremen.de
Lynn Schriml	Assistant Professor	University of Maryland School of Medicine	lschriml@som.umaryland.edu
Erica Sodergren	Assistant Director, Research Associate Professor of Genetics	Washington University in Saint Louis School of Medicine	esodergr@genome.wustl.edu
Peter Sterk	Staff member	University of Oxford e-Research Centre	psterk1@gmail.com
Steven Stones-Havas	Senior Consultant	Biomatters Limited	brett@biomatters.com
Granger Sutton	Professor	J. Craig Venter Institute	gsutton@jcvi.org
Tatiana Tatusova	Staff Scientist	NCBI NLM NIH	tatiana@ncbi.nlm.nih.gov
Petra Ten Hoopen	scientific database curator	ENA	petra@ebi.ac.uk

Meeting participants (cont.)

Full Name (First Middle Last)	Job Title	CompanyOrganization	Email
Michael Trizna	Data Management Specialist	Smithsonian	triznam@si.edu
David W Ussery	professor	CBS, Dept. Systems Biology, DTU	dave@cbs.dtu.dk
Ramona Walls	Scientific Analyst	The iPlant Collaborative	rwalls@iplantcollaborative.org
Qiong Wang	IT	MSU	wanqgion@msu.edu
Spencer Wells	Director	The Genographic Project	spwells@ngs.org
Trish Whetzel	Outreach Coordinator	NCBO/Stanford University	whetzel@stanford.edu
Owen White	Professor, Associate Director	Institute for Genome Sciences	owhite@som.umaryland.edu
Andreas Wilke	Bioinformatics Software Specialist	Argonne National Laboratory	wilke@mcs.anl.gov
K Eric Wommack	Professor	Univ of Delaware	wommack@dbi.udel.edu
Alison Yao	Program Officer	NIAID	yaoal@niaid.nih.gov
Pelin Yilmaz	Postdoctoral researcher	Microbial Genomics Group	pyilmaz@mpi-bremen.de
Justin Zook	Biomedical Engineer	National Institute of Standards and Technology	jzook@nist.gov